Welcome from the Scientific Program Committee Chair

On behalf of the 2015 Scientific Program Committee, I am welcoming you—informatics researchers, practitioners, professionals, students, trainees, colleagues, and most importantly friends—to the AMIA 2015 Annual Symposium here in San Francisco! Hosting AMIA 2015 on the West Coast near Silicon Valley, we offer a special warm welcome to our colleagues in computer science, information science, computational biology, social science. You are joining colleagues from translational bioinformatics, clinical research informatics, clinical informatics, consumer health informatics, and public health informatics eager to meet, collaborate, and learn.

Fueled by the Power of Informatics, the AMIA Symposium is the premier event to present and learn about the latest scientific discoveries, join your colleagues and peers in networking and fellowship, begin lasting collaboration and professional friendships, engage in spirited discussions and debates on controversial informatics topics, learn from others in your field, and connect to the mentors and leaders in informatics.

With two tracks in Foundations and Applied Informatics—often interwoven through joint sessions that allow the leap from base pairs to bench to bedside to patients to populations and back around—the symposium offers a home for all informatics experts to present and share their experience and simultaneously learn both about implementation as well as research.

AMIA 2015 sessions feature full-length papers, abstract presentations, and posters, panel discussions including didactic and interactive panels with extensive audience participation, cutting-edge policy presentations, keynotes, tutorials, system demonstrations, and pre-Symposium workshops. Please review the Schedule-at-a-Glance and use our brand-new mobile App, which is an excellent tool for getting the most out of the meeting.

Highlights

- Keynote Speakers: Avi Rubin, PhD, is a Professor of Computer Science at Johns Hopkins University and will address Information Security in his talk “Hacking Health Care”. Robert M. Wachter, MD is Professor and Associate Chairman of the Department of Medicine at the University of California, San Francisco will focus on Patient Safety
- New in 2015, the Connect-a-Thon pairs AMIA members with leaders in the health care start-up companies to solve problems and develop lasting collaborations.
- Learning Showcase presents new research-focused sessions in the Exhibition Hall
- Student Design Challenge provides students or teams of students with the opportunity to design a solution to an informatics problem.
- The High School Scholars Competition will bring again fresh faces to AMIA to present their work in biomedical informatics summer programs across the country.

I would like to thank my Scientific Program Committee and particularly Vice Chair for Foundations Riccardo Bellazzi, Vice Chair for Applications Sarah Ingersoll and Wanda Pratt, the AMIA 2016 SPC Chair. Along with the amazing AMIA staff, we orchestrated the planning, peer-review, and program design for the AMIA 2015 Annual Symposium. Last, but certainly not least, I would like to thank the 600-plus informaticians who served as subject matter experts in our peer-review process. The success of the Symposium program truly begins with them.

We believe the Scientific Program Committee has prepared a fantastic program. Please be sure to avail yourself of all that the Symposium has to offer—or as much as you can fit in!

Christoph U. Lehmann, MD, FAAP, FACMI
Chair, AMIA 2015 Scientific Program Committee
@culehmann
2015 AMIA Board of Directors

Officers

Blackford Middleton, MD, MPH, MSc, FACMI
Chair
2014 - 2015

Thomas H. Payne, MD
University of Washington
Chair-elect
2015

Steven Labkoff, MD, FACP
Intelligent Medical Objects (IMO)
Treasurer
2014 - 2015

Eneida A Mendonça, MD, PhD, FACMI, FAAP
University of Wisconsin – Madison
Secretary
2014 – 2015

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

Directors

Martha Bennett Adams, MA, MD, FACP
Working Group Steering Committee Chair
Duke University School of Medicine
2013 – 2015

Wendy W. Chapman, PhD, FACMI
University of Utah
2014 - 2016

Theresa Cullen, MD, MS
Veterans Health Administration
2015-2017

Patricia Dykes, DNSc, MA, RN, FACMI
Brigham and Women’s Hospital
2014 – 2016

Peter J. Embi, MD, MS, FACMI
The Ohio State University
2015 - 2017

Cynthia S. Gadd, PhD, FACMI
Vanderbilt University
2015-2017

John H. Holmes, PhD, FACMI
University of Pennsylvania
2013 - 2015

Curtis P. Langlotz, MD, PhD, FACMI, FSIIM
Stanford University
2015 - 2016

Neil Sarkar, PhD, FACMI
Brown University
2014 - 2016

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

Justin B. Starren, MD, PhD, FACMI
Northwestern University Biomedical Informatics Center
2013 - 2015

Jessica Tenenbaum, PhD
Duke Translational Medicine Institute
2015 - 2017

Ex-Officio Board Members

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

Michael Kamerick, BSCS
Industry Advisory Council Chair
TriNetx, Inc.
2015

Tiffany Kelley, PhD, MBA, RN
Student Working Group Representative
2014 to 2015
Philip Kroth, MD, MS
Academic Forum Executive Committee Chair
University of New Mexico Health Sciences Library and Informatics Center
2015
AMIA 2015 Scientific Program Committee

Christoph Lehmann, MD, FAAP, FACMI
Vanderbilt University
Chair

Riccardo Bellazzi, PhD
Università di Pavia
Foundations Track Chair

Sarah Ingersoll, RN, MBA
University of Southern California
Applications Track Chair

Jos Aarts, PhD, FACMI
University at Buffalo

Martha Adams, MD, MA,
Duke University Medical Center

Swati Chakraborty, MS
NYU Langone Medical Center

James Cimino, MD
University of Alabama at Birmingham

Amar Das, MD, PhD
Dartmouth College

Jason Doctor, PhD
University of Southern California

Henry Feldman, MD
Beth Israel Deaconess Medical Center

Lewis Frey, PhD
Medical University of South Carolina

Kate Fultz Hollis, MS
OHSU

Adi Gundlapalli, MD, PhD, MS
University of Utah School of Medicine

Milos Hauskrecht, PhD
University of Pittsburgh
Andre Kushniruk, PhD
University of Victoria

Albert Lai, PhD
The Ohio State University

Brandyn Lau
Johns Hopkins University

Tze-Yun Leong, PhD
National University of Singapore, School of Computing

Olena Mamykina, PhD
Columbia University

Michael Matheny, MD, MS, MPH
Vanderbilt University Medical Center

Julie McGowan, PhD

Daniella Meeker
RAND Corporation

Shawn Murphy, MD, PhD, FACMI
Massachusetts General Hospital

Joyce Niland, PhD
City of Hope National Medical Center

Carolyn Petersen, MBI, MS
Mayo Clinic

Wanda M. Pratt, PhD
University of Washington
AMIA 2016 Chair

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

S. Trent Rosenbloom, MD, MPH
Vanderbilt University Medical Center Dept of Biomedical Informatics

Lipika Samal, MD
Brigham and Women's Hospital

Charlotte A. Seckman, PhD, RN, BC, MSN
University of Maryland Baltimore
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aakre, Christopher</td>
<td>Albaghdadi, Mazen</td>
<td>Anton, Bonnie</td>
</tr>
<tr>
<td>Abbaszadeghan, Hamed</td>
<td>Albers, David</td>
<td>Antonille, Tanya Santelli</td>
</tr>
<tr>
<td>Abbott, Amy Ann</td>
<td>Aldridge, Deborah Ann</td>
<td>Aphinyanaphongs, Yindalon</td>
</tr>
<tr>
<td>Abdelrahman, samir E</td>
<td>Alexander, Gretchen B.</td>
<td>Ariosto, Deborah</td>
</tr>
<tr>
<td>Abedtash, Hamed</td>
<td>Alexander, Alex</td>
<td>Arnold, Corey W</td>
</tr>
<tr>
<td>Abend, Aaron H</td>
<td>Alexander, Gregory L</td>
<td>Aronson, Alan R</td>
</tr>
<tr>
<td>Abraham, Ivo</td>
<td>AlHuwait, Dari</td>
<td>Artz, Nathan</td>
</tr>
<tr>
<td>AbuSalah, Ahmad M</td>
<td>Alipanah, Neda</td>
<td>Asaro, Phillip</td>
</tr>
<tr>
<td>Acharya, Amit</td>
<td>Allin, Sonya</td>
<td>Ash, Nachman</td>
</tr>
<tr>
<td>Adam, Terrence J</td>
<td>Alloni, Anna</td>
<td>Ash, Joan S</td>
</tr>
<tr>
<td>Adams, Martha B</td>
<td>Almeida, Jonas S</td>
<td>Ashfaq, Shazia</td>
</tr>
<tr>
<td>Adams, Samantha</td>
<td>Alonso-Calvo, Raul</td>
<td>Ashish, Naveen</td>
</tr>
<tr>
<td>Adelman, Jason</td>
<td>Alsip, Jorge A</td>
<td>Aslani Dr., Azam</td>
</tr>
<tr>
<td>Agrawal, Ankur</td>
<td>Altamore, Rita</td>
<td>Au, Alan P.</td>
</tr>
<tr>
<td>Ahmad, Suzan</td>
<td>Amato, Mary G</td>
<td>AU, Daisy Yat See</td>
</tr>
<tr>
<td>Ahmed, Waled Amed</td>
<td>Ambert, Kyle H.</td>
<td>Augestad, Knut Magne</td>
</tr>
<tr>
<td>Ahmed, Adil</td>
<td>Amith, Muhammad Faheem</td>
<td>Avali, Viji</td>
</tr>
<tr>
<td>Ahmed, Zeeshan</td>
<td>Ammenwerth, Elske</td>
<td>Avrunin, George S.</td>
</tr>
<tr>
<td>Ahuja, Monika</td>
<td>Anand, Vibha</td>
<td>Ayres, Elaine</td>
</tr>
<tr>
<td>Ahumada, Luis</td>
<td>Ancker, Jessica S</td>
<td>Ayubi, Soleh U</td>
</tr>
<tr>
<td>Airan-Javia, Subha L</td>
<td>Anders, Shilo H</td>
<td>Azadmanjir, Zahra</td>
</tr>
<tr>
<td>Akmaev, Viatcheslav</td>
<td>Andrus, Charles H</td>
<td>Babic, Ankica</td>
</tr>
<tr>
<td>Alafaireet, Patricia E</td>
<td>Angier, Heather</td>
<td>Backonja, Uba</td>
</tr>
<tr>
<td>Alaiad, Ahmad</td>
<td>Annechiarico, Robert P</td>
<td>Badger, Martha Kimpton</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-------------------------------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Bagayoko, Cheikh O</td>
<td>Bell, Heidi G</td>
<td>Blecker, Saul</td>
</tr>
<tr>
<td>Bai, Tian</td>
<td>Ben Souissi, Souhir</td>
<td>Bleimeyer, Robert</td>
</tr>
<tr>
<td>Bailey, James Michael</td>
<td>Benavides-Torres, Raquel Alicia</td>
<td>Blondon, Katherine</td>
</tr>
<tr>
<td>Bakken, Suzanne</td>
<td>Benham-Hutchins, Marge M</td>
<td>Bodenreider, Olivier</td>
</tr>
<tr>
<td>Baldwin, Kevin M</td>
<td>Benitez, Sonia Elizabeth</td>
<td>Boland, Mary Regina</td>
</tr>
<tr>
<td>Banerjee, Ashis Gopal</td>
<td>Benuzillo, Jose</td>
<td>Bona, Jonathan P</td>
</tr>
<tr>
<td>Banger, Alison</td>
<td>Bergamasco, Leila</td>
<td>Boone, Keith W</td>
</tr>
<tr>
<td>Banning, Pamela</td>
<td>Berkovich, Barbara</td>
<td>Booth, James</td>
</tr>
<tr>
<td>Bansal, Arvind</td>
<td>Berner, Eta S.</td>
<td>Booth, David</td>
</tr>
<tr>
<td>Baptista, Roberto Silva</td>
<td>Berntsen MD, Dr. Med, Gro</td>
<td>Borbolla, Damian Alberto</td>
</tr>
<tr>
<td>Barnes, Laura</td>
<td>Berrios, Daniel C</td>
<td>Boren, Suzanne A</td>
</tr>
<tr>
<td>Barone, Eleanor Jane</td>
<td>Berry, Donna</td>
<td>Borland PhD, David</td>
</tr>
<tr>
<td>Barros, Roberto Souto Maior de</td>
<td>Berry, Andrew</td>
<td>Borlawsky, Tara</td>
</tr>
<tr>
<td>Bashyam, Vijayaraghavan</td>
<td>Berster, Bjoern-Toby</td>
<td>Botts, Nathan</td>
</tr>
<tr>
<td>Basu Roy, Senjuti</td>
<td>Bhatia, Haresh</td>
<td>Bouaud, Jacques</td>
</tr>
<tr>
<td>Baumgart, Leigh A</td>
<td>Bhattacharya, Sanmitra</td>
<td>Bouayad, Lina</td>
</tr>
<tr>
<td>bayraktar, Mustafa</td>
<td>Bhuvaneshwar, Krithika</td>
<td>Boughorbel, Sabri</td>
</tr>
<tr>
<td>Beard, Rachel</td>
<td>Bian, Jiang</td>
<td>Bouras, Adam</td>
</tr>
<tr>
<td>Bechtold, Thomas P</td>
<td>Bickel, Jonathan</td>
<td>Bourgeois, Florence</td>
</tr>
<tr>
<td>Bedra, McKenzie</td>
<td>Bickmore, Timothy</td>
<td>Boussadi, Abdelali</td>
</tr>
<tr>
<td>Beeler, Patrick E</td>
<td>Binkheder, Samar</td>
<td>Bowles, Kathryn H</td>
</tr>
<tr>
<td>Beer, Idal</td>
<td>Bjarnadottir, Ragnhildur I.</td>
<td>Boyce, Richard David</td>
</tr>
<tr>
<td>Beidas, Sary Omar</td>
<td>Blake, Catherine</td>
<td>Boyd, Andrew</td>
</tr>
<tr>
<td>Bejan, Cosmin</td>
<td>Blaz, Jacquelyn Williamson</td>
<td>Boyd, Angela</td>
</tr>
</tbody>
</table>
AMIA 2015 REVIEWERS

Bradford, Jennifer
Bradley, David James
Brandão, Carolina Soares
Brasil Pintarelli, Gulherme
Braunstein, Mark L
Breene, Michael
Breitenstein, Matthew K.
Brennan, Caitlin W
Brennan, Patricia F
Brenner, Samantha K.
Bressan, Nadja
Bright, Tiffani J
Brixey, Juliana J
Brochhausen, Mathias
Broens, Tom
Bronsburg Mr., Steve Edward
Brown, Laura E
Brown III, William
Browning, Doug
Buckeridge, David
Bucur, Anca
Buis, Lorraine
Burckhardt, Philipp
Burke, Harry B.
Bustamante, Claudia
Butcher, Ryan
Butler, Jorie Michaela
Butler, Keith A
Byrd, Roy J
Byrne, Colene
Caballero Barajas, Karla Lizbeth
Caban, Jesus J
Cahan, Amos
Cai, Binghuang
Cai, Shu
Cairelli, Michael J.
Callahan, Alison
Callahan, Tiffany Jean Lee
Calvo, Richard Y
Cameron, Joshua
Campbell, James Richard
Campbell, Robert C
Campbell, Walter Scott
Campbell, Terrance R
Campion Jr, Thomas Richmond
Campos, Manuel
Cantor, Michael
Cao, Hui
Caraballo, Pedro J
Cardillo, Elena
Carrell, David S
Carroll, Robert J
Carson, Kimberly J
Carson, Matthew
Carter, Ernest
Carvalho Jr., Marcelo Antonio de
Casper, Gail R.
Castellanos, Arturo
Castro, Victor M
Catalina, Danis
Cato, Kenrick
Cazeneuve, Herve
Ceusters, Werner
Chall, Steven Alan
Chambers, Nadene
Champagne, Tiffany
Chandran, Uma
Chang, Yu-Hui
Chapman, Elizabeth S
Charney, Pam
Chartash, David
Chary, Michael
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chase, Dian A.</td>
<td>Choquet, Remy</td>
<td>Corley, Sarah</td>
</tr>
<tr>
<td>Chau, Ngan</td>
<td>Chou, David</td>
<td>Correia, André</td>
</tr>
<tr>
<td>Chaudhry, Rajeev</td>
<td>Chung, Jane</td>
<td>Costa Monteiro, Lucia Maria</td>
</tr>
<tr>
<td>Chen, Xue-Wen</td>
<td>Chung, Arlene E.</td>
<td>Costello, Kaitlin</td>
</tr>
<tr>
<td>Chen, You</td>
<td>Cimino, Christopher</td>
<td>Courtney, Paul Kenneth</td>
</tr>
<tr>
<td>Chen, Yukun</td>
<td>Cimino, James J</td>
<td>Courtney, Karen L</td>
</tr>
<tr>
<td>Chen, Lujie</td>
<td>Clark, Cheryl</td>
<td>Courtot, Melanie</td>
</tr>
<tr>
<td>Chen, Weiheng</td>
<td>Claypool, Stephen</td>
<td>Crabtree, Elizabeth A.</td>
</tr>
<tr>
<td>Chen, Jonathan H</td>
<td>Clements, Marsha</td>
<td>Craven, Catherine K</td>
</tr>
<tr>
<td>Chen, Robert</td>
<td>Clermont, Gilles</td>
<td>Crew, Katherine</td>
</tr>
<tr>
<td>Chen, Elizabeth S.</td>
<td>Co Jr., Manuel C.</td>
<td>Cronin, Robert M</td>
</tr>
<tr>
<td>Cheng, Eric</td>
<td>Cohen, Genna Rebecca</td>
<td>Cronkite, David J</td>
</tr>
<tr>
<td>Cheng, Yu</td>
<td>Cohen, Trevor</td>
<td>Crooks, John A</td>
</tr>
<tr>
<td>Cheng, Alex</td>
<td>Cole, Curtis L</td>
<td>Crotty, Bradley H</td>
</tr>
<tr>
<td>Cheon, Eunjeong</td>
<td>Cole, Elodia</td>
<td>Cui, Licong</td>
</tr>
<tr>
<td>Chi, Chih-Lin</td>
<td>Cole-Lewis, Heather</td>
<td>Cullen, Theresa</td>
</tr>
<tr>
<td>Chiang, Michael F.</td>
<td>Collins, Sarah</td>
<td>Culley, Joan Marie</td>
</tr>
<tr>
<td>Chiaravalloti, Maria Teresa</td>
<td>Coloma, Preciosa M</td>
<td>Cummins, Mollie R</td>
</tr>
<tr>
<td>Chih, Ming-Yuan</td>
<td>Connolly, John J.</td>
<td>Curcin, Vasa</td>
</tr>
<tr>
<td>Chin, Si-Chi</td>
<td>Connolly, Brian</td>
<td>Curioso, Walter H</td>
</tr>
<tr>
<td>Cho, Insook</td>
<td>Connolly, Natalia</td>
<td>Curtis, Dorothy W</td>
</tr>
<tr>
<td>Cho, InSook</td>
<td>Constantin, Alexandra</td>
<td>Da Silveira, Marcos</td>
</tr>
<tr>
<td>Cho, Jason H.D.</td>
<td>Conway, Mike</td>
<td>Dagliati, Arianna</td>
</tr>
<tr>
<td>Choi, Jeeyae</td>
<td>Cooley, Sarah</td>
<td>Dai, Peng</td>
</tr>
<tr>
<td>AMIA 2015 REVIEWERS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dai, Hong-Jie</td>
<td>Demner-Fushman, Dina</td>
<td></td>
</tr>
<tr>
<td>Dalianis, Hercules K</td>
<td>DeMuro Esq., Paul Robert</td>
<td></td>
</tr>
<tr>
<td>Dalrymple, Prudence</td>
<td>Denaxas, Spiros</td>
<td></td>
</tr>
<tr>
<td>Daneshvari, Shamsi Rebecca</td>
<td>Denny, Joshua C</td>
<td></td>
</tr>
<tr>
<td>Danielsson-Ojala, Riitta</td>
<td>Dery, Samuel K K</td>
<td></td>
</tr>
<tr>
<td>Dantes, Raymund</td>
<td>Devarakonda, Murthy V</td>
<td></td>
</tr>
<tr>
<td>Daphtary, Kshama</td>
<td>Dew, Kristin N</td>
<td></td>
</tr>
<tr>
<td>Darcy, Niamh</td>
<td>Dewan, Peehoo</td>
<td></td>
</tr>
<tr>
<td>Das, Anirudha</td>
<td>Dexheimer, Judith</td>
<td></td>
</tr>
<tr>
<td>Dasari, Surendra</td>
<td>Dey, Sanjoy</td>
<td></td>
</tr>
<tr>
<td>Daskivich, Lauren Patty</td>
<td>Dhombres, Ferdinand</td>
<td></td>
</tr>
<tr>
<td>Davis, Jesse</td>
<td>Dhopheshwarkar, Neil</td>
<td></td>
</tr>
<tr>
<td>Davis, Selena</td>
<td>Diallo, Gayo</td>
<td></td>
</tr>
<tr>
<td>Davis, Sharon E</td>
<td>Diaz, Martín Miguel</td>
<td></td>
</tr>
<tr>
<td>Dcruz, Jina J</td>
<td>Dinakarpandian, Deendayal</td>
<td></td>
</tr>
<tr>
<td>de Bruijn, Berry</td>
<td>Divita, Guy</td>
<td></td>
</tr>
<tr>
<td>de Sisternes, Luis</td>
<td>Dixon, Brian E</td>
<td></td>
</tr>
<tr>
<td>deBronkart, Dave</td>
<td>Dixon-Lee, Claire</td>
<td></td>
</tr>
<tr>
<td>Degoulet, Patrice</td>
<td>Dligach, Dmitriy</td>
<td></td>
</tr>
<tr>
<td>Dekker, Andre</td>
<td>Do, Nhan V</td>
<td></td>
</tr>
<tr>
<td>Del Fiol, Guilherme</td>
<td>Do, Nhan V</td>
<td></td>
</tr>
<tr>
<td>Deleger, Louise</td>
<td>Doan, Son</td>
<td></td>
</tr>
<tr>
<td>Delen, Dursun</td>
<td>Doberne, Julie W.</td>
<td></td>
</tr>
<tr>
<td>Demiris, George</td>
<td>Dobson, Richard James Butler</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doctor, Jason N</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doing-Harris, Kristina Mary</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doshi, Riddhi</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Douglas, Gerald P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dowding, Dawn</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doyle, Glynda</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doyon, Katherine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Driscoll, Patricia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Du, Jingcheng</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dua, Prerna</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duclos, Catherine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dufendach, Kevin R</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duke, Jon</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dullabh, Prashila</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dumontier, Michel</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duncan, Jeffrey</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dunn Lopez, Karen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dutra, Ines</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dzomba, Bari</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Edinger, Tracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Edmunds, Margo</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Effken, Judith</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Eguale, Tewodros</td>
<td>Fearn, Paul</td>
<td>Fontelo, Paul</td>
</tr>
<tr>
<td>Eichenhorn, Michael</td>
<td>Febretti, Alessandro</td>
<td>Foraker, Randi E</td>
</tr>
<tr>
<td>Eisenstein, Eric</td>
<td>Fedorov, Andriy</td>
<td>Fort, Daniel</td>
</tr>
<tr>
<td>Ekberg, Joakim</td>
<td>Felmeister, Alex S</td>
<td>Fouliis, Philip</td>
</tr>
<tr>
<td>Eldredge, Christina</td>
<td>Feng, Yuanyuan</td>
<td>Franklin Dr., Amy</td>
</tr>
<tr>
<td>Elhadad, Noemie</td>
<td>Feng, Mengling</td>
<td>Franzke, Laura H</td>
</tr>
<tr>
<td>El-Kareh, Robert</td>
<td>Fenton, Susan H</td>
<td>Freimuth, Robert R</td>
</tr>
<tr>
<td>Engen-Verheul, Mariette van</td>
<td>Fernandez-Breis, Jesualdo</td>
<td>Friedman, Carol</td>
</tr>
<tr>
<td>Epps, Mika</td>
<td>Fernandez-Luque PhD, Luis</td>
<td>Frisse, Mark E</td>
</tr>
<tr>
<td>Eriksson, Henrik</td>
<td>Ferraro, Jeffrey P</td>
<td>Fronza, Carlos F</td>
</tr>
<tr>
<td>Eschler, Jordan</td>
<td>Figueroa, Rosa Liliana</td>
<td>Fuji, Kevin T.</td>
</tr>
<tr>
<td>Ethier, Jean-Francois</td>
<td>Finch, Dezon</td>
<td>Fung, Kin Wah</td>
</tr>
<tr>
<td>Evans, Scott</td>
<td>Finnell II, John</td>
<td>Furman, Amy</td>
</tr>
<tr>
<td>Evans, R. Scott</td>
<td>Fischer, Shira H</td>
<td>Futoma, Joseph</td>
</tr>
<tr>
<td>Evans, Heather L</td>
<td>Fisher, Arielle</td>
<td>Gabetta, Matteo</td>
</tr>
<tr>
<td>Ewing, Danita Lee</td>
<td>Fiszman, Marcelo</td>
<td>Gadabu, Oliver</td>
</tr>
<tr>
<td>Faiola, Anthony</td>
<td>FitzHenry, Fern</td>
<td>Gagner, Jason</td>
</tr>
<tr>
<td>Fan, Jung-wei</td>
<td>Flegel, Willy A</td>
<td>Gakuba, Richard</td>
</tr>
<tr>
<td>Fan, Jing</td>
<td>Fleisher, Linda</td>
<td>Gallo De Moraes, Alice</td>
</tr>
<tr>
<td>Farouq, Faisal</td>
<td>Fleming, Karol L</td>
<td>Galvez, Jorge A.</td>
</tr>
<tr>
<td>Farri, Oladimeji</td>
<td>Florez-Arango, Jose Fernando</td>
<td>Gamache, Roland</td>
</tr>
<tr>
<td>Farrugia, Sherry N</td>
<td>Flynn, Allen J</td>
<td>Gandhi, Purav</td>
</tr>
<tr>
<td>Faysel, Mohammad</td>
<td>Fodeh, Samah Jamal</td>
<td>Ganesan, Kavita</td>
</tr>
<tr>
<td>Faysel, Mohammad</td>
<td>Fodeh, Samah Jamal</td>
<td>Gao, Grace</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-------------------------------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Garcia Arguello, Lisbeth Y</td>
<td>Goel, Anupam</td>
<td>Greenberg, Jeffrey O</td>
</tr>
<tr>
<td>Gardeux, Vincent R. J.</td>
<td>Goel, Satyender</td>
<td>Gregory, Judith</td>
</tr>
<tr>
<td>Gartner, Daniel</td>
<td>Goggins, Kathryn</td>
<td>Gregson, Cassie</td>
</tr>
<tr>
<td>Gartrell, Kyungsook</td>
<td>Gold, Jeffrey A</td>
<td>Grinspan, Zachary M</td>
</tr>
<tr>
<td>Garvin, Jennifer Hornung</td>
<td>Gold, Rachel</td>
<td>Grose, Tim</td>
</tr>
<tr>
<td>Gatz, Jennifer</td>
<td>Goldschmidt, Yaara</td>
<td>Grosjean, Julien</td>
</tr>
<tr>
<td>Gaynor, Mark</td>
<td>Goldspiel, Barry R</td>
<td>Grouin, Cyril</td>
</tr>
<tr>
<td>Ge, Yaorong</td>
<td>Goldstein, Mary K.</td>
<td>Ground, Anita Rose</td>
</tr>
<tr>
<td>Gee, Perry M</td>
<td>Gonçalves, Rafael S.</td>
<td>Groza, Tudor</td>
</tr>
<tr>
<td>Gephart, Sheila M</td>
<td>Gong, Yang</td>
<td>Grundmeier, Robert</td>
</tr>
<tr>
<td>Gerhard, Glenn</td>
<td>Gonzalez, Nathan C</td>
<td>Guappone, Kenneth P</td>
</tr>
<tr>
<td>Gerkovich, Mary M</td>
<td>González Ferrer, Arturo</td>
<td>Guida, Kari</td>
</tr>
<tr>
<td>Gharehbaghi, Arash</td>
<td>Goodwin, Rebecca Marci</td>
<td>Guidry, Alicia</td>
</tr>
<tr>
<td>Ghassemi, Mohammad Mahdi</td>
<td>Goonawardene, Nadee</td>
<td>Guimarães, Marcelo Paiva</td>
</tr>
<tr>
<td>Ghebremariam, Yohannes T</td>
<td>Gordon, William</td>
<td>Gundlapalli, Adi V</td>
</tr>
<tr>
<td>Giang, Phan</td>
<td>Gottesman, Omri</td>
<td>Gurupur, Varadraj Prabhu</td>
</tr>
<tr>
<td>Gibson, Bryan</td>
<td>Gotz, David</td>
<td>Gururaj, Anupama Edward</td>
</tr>
<tr>
<td>Gilbertson, David G</td>
<td>Gourab, Krishnaj</td>
<td>Gwadry-Sridhar, Femida</td>
</tr>
<tr>
<td>Gilder, Jason</td>
<td>Govan-Jenkins, Wanda Yyonnette</td>
<td>Haas, Stephanie W</td>
</tr>
<tr>
<td>Gimenez, Francisco</td>
<td>Goyal, Dev</td>
<td>Hack, Lori</td>
</tr>
<tr>
<td>Giunti, Guido</td>
<td>Grabar, Natalia</td>
<td>Haddad, Peter</td>
</tr>
<tr>
<td>Giussi, Maria Victoria</td>
<td>Grando, Maria Adela</td>
<td>Hadji, Brahim</td>
</tr>
<tr>
<td>Glaser, John</td>
<td>Grannis, Shaun</td>
<td>Hadley, Dexter</td>
</tr>
<tr>
<td>Gobbel, Glenn T</td>
<td>Grasso, Michael A.</td>
<td>Hafeez, Baria</td>
</tr>
<tr>
<td>Hagen, Michael D.</td>
<td>Hausam, Robert</td>
<td>Ho, Kendall</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------</td>
<td>------------</td>
</tr>
<tr>
<td>Haggstrom, David</td>
<td>Hawkins, Jared B.</td>
<td>Ho, Samuel B.</td>
</tr>
<tr>
<td>Hajihashemi, Zahra</td>
<td>Hays, David</td>
<td>Ho, King Chung</td>
</tr>
<tr>
<td>Hales, Joseph W.</td>
<td>Hazen, Rebecca J</td>
<td>Ho, Ping Chong</td>
</tr>
<tr>
<td>Hall, Susan</td>
<td>He, Ze</td>
<td>Hochheiser, Harry</td>
</tr>
<tr>
<td>Hall, Eric S</td>
<td>He, Zhe</td>
<td>Hoffman, James M</td>
</tr>
<tr>
<td>Hall, Amanda K</td>
<td>He, Yongqun</td>
<td>Hoffman, Mark</td>
</tr>
<tr>
<td>Hameed, Farhan</td>
<td>He, Shan</td>
<td>Hogan, William</td>
</tr>
<tr>
<td>Hammadi, Slim</td>
<td>Heale, Bret S.E.</td>
<td>Hogarth, Mike</td>
</tr>
<tr>
<td>Hammond, Kenric W</td>
<td>Heard, Kevin M</td>
<td>Hoh, Tim</td>
</tr>
<tr>
<td>Hammond, William Ed</td>
<td>Hebert, Courtney Lee</td>
<td>Hollberg, Julie</td>
</tr>
<tr>
<td>Hamon, Thierry</td>
<td>Heinrich, Julia N</td>
<td>Holmes, John H.</td>
</tr>
<tr>
<td>Han, Eungyoung</td>
<td>Hekmatnejad, Mohammad</td>
<td>Holt, Jonathan</td>
</tr>
<tr>
<td>Han, Simon X</td>
<td>Hempelmann, Christian F.</td>
<td>Homer, Mark L</td>
</tr>
<tr>
<td>Hanauer, David</td>
<td>Henao, Ricardo</td>
<td>Hook, Mary L.</td>
</tr>
<tr>
<td>Hanauer, David A</td>
<td>Henriksson, Aron</td>
<td>Hornstein, Rachel</td>
</tr>
<tr>
<td>Hanna, Josh</td>
<td>Herasevich, Vitaly</td>
<td>Horridge, Matthew</td>
</tr>
<tr>
<td>Hao, Tianyong</td>
<td>Hermes-DeSantis, Evelyn</td>
<td>Horsky, Jan</td>
</tr>
<tr>
<td>Harle, Christopher A</td>
<td>Herout, Jennifer</td>
<td>Hosomura, Naoshi</td>
</tr>
<tr>
<td>Harrington, Linda</td>
<td>Hettinger, Aaron Zachary</td>
<td>Hosseini, Masoud</td>
</tr>
<tr>
<td>Harrington, Craig</td>
<td>Hickman, Timothy P</td>
<td>Hovenga, Evelyn JS</td>
</tr>
<tr>
<td>Hartzler, Andrea L.</td>
<td>Hickman, Thu-Trang</td>
<td>Hoyt, Robert E</td>
</tr>
<tr>
<td>Hasan, Omar</td>
<td>Hillman, JoAnna L.</td>
<td>Hribar, Michelle R</td>
</tr>
<tr>
<td>Hassanpour, Saeed</td>
<td>Hirsch, Annemarie Gregory</td>
<td>Hruby, Gregory William</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Hsu, Chun-Nan</td>
<td>Jacobs, Maia</td>
<td>Johnson, Todd R</td>
</tr>
<tr>
<td>Hsu, William</td>
<td>Jacobs, Robin J.</td>
<td>Johnston, Jessica</td>
</tr>
<tr>
<td>Hsueh, Pei-Yun Sabrina</td>
<td>Jadhav, Ashutosh</td>
<td>Johnston, Douglas</td>
</tr>
<tr>
<td>Hu, Zhen</td>
<td>Jaffe, Charles</td>
<td>Jones, Josette F.</td>
</tr>
<tr>
<td>Hu, Lu</td>
<td>Jagannathan, Vasudevan</td>
<td>Jones-Diette, Julie Susan</td>
</tr>
<tr>
<td>Hu, Jianying</td>
<td>Jain, Nilesh</td>
<td>Jouhet, Vianney</td>
</tr>
<tr>
<td>Huang, Zhengxing</td>
<td>Jain, Viral</td>
<td>Juhra, Christian</td>
</tr>
<tr>
<td>Hudak, Christine Angela</td>
<td>Jain, Anil</td>
<td>Jung, Changmi</td>
</tr>
<tr>
<td>Huffman, Keith</td>
<td>Janeja, Vandana</td>
<td>Jung, Hyunggu</td>
</tr>
<tr>
<td>Hughes, Kevin S</td>
<td>Jao, Chiang</td>
<td>Juster, Iver A</td>
</tr>
<tr>
<td>Huh, Jina</td>
<td>Jenders, Robert Allen</td>
<td>Kaelber, David</td>
</tr>
<tr>
<td>Hull, Susan</td>
<td>Jeong, In Cheol</td>
<td>Kahler, Zachary Piazza</td>
</tr>
<tr>
<td>Hulse, Nathan C</td>
<td>Ji, Xiaonan</td>
<td>Kahn Jr, Charles E.</td>
</tr>
<tr>
<td>Husband, Andrew K</td>
<td>Ji, Xiang</td>
<td>Kakkar, Reshma</td>
</tr>
<tr>
<td>Huser, Vojtech</td>
<td>Jiang, Min</td>
<td>Kale, David</td>
</tr>
<tr>
<td>Hussein, Rada</td>
<td>Jiang, Guoqian</td>
<td>Kalenderian, Elsbeth</td>
</tr>
<tr>
<td>Hynes, Denise M.</td>
<td>Jiang, Silis Y.</td>
<td>Kalpathy-Cramer, Jayashree</td>
</tr>
<tr>
<td>Hyun, Sookyung</td>
<td>Jiang, Junhui</td>
<td>Kalton, Alan</td>
</tr>
<tr>
<td>Imler, Timothy David</td>
<td>Jimeno Yepes, Antonio</td>
<td>Kamaauu, Aaron</td>
</tr>
<tr>
<td>Iranmehr, Arya</td>
<td>Jing, Xia</td>
<td>Kandula, Vijayabhaskar Reddy</td>
</tr>
<tr>
<td>Iribarren, Sarah</td>
<td>Johnson, Craig William</td>
<td>Kang, Hong</td>
</tr>
<tr>
<td>Ito, Márcia</td>
<td>Johnson, Sharon Ann</td>
<td>Kang, Tian</td>
</tr>
<tr>
<td>Ivory, Catherine H</td>
<td>Johnson, Constance M</td>
<td></td>
</tr>
<tr>
<td>Jackson, Gretchen Purcell</td>
<td>Johnson, Stephen Bennett</td>
<td></td>
</tr>
<tr>
<td>Reviewer Name</td>
<td>Reviewer Name</td>
<td>Reviewer Name</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Kanter, Andrew Stuart</td>
<td>Khodambashi, Soudabeh</td>
<td>Kohn, Martin S</td>
</tr>
<tr>
<td>Kanthawala, Shaheen</td>
<td>Kilicoglu, Halil</td>
<td>Kommera, Naveen R</td>
</tr>
<tr>
<td>Kapoor, Akshat</td>
<td>Kim, Christopher Y</td>
<td>Kondo, Yoshiaki</td>
</tr>
<tr>
<td>Karami, Amir</td>
<td>Kim, Ju Han</td>
<td>Kooienga, Sarah A.</td>
</tr>
<tr>
<td>Karavite, Dean J</td>
<td>Kim, Tae Youn</td>
<td>Koola, Jejo David</td>
</tr>
<tr>
<td>Kartoun, Uri</td>
<td>Kim, Sun</td>
<td>Korkontzelos, Ioannis</td>
</tr>
<tr>
<td>Kashani, Kianoush</td>
<td>Kim, Won</td>
<td>Koru, Gunes</td>
</tr>
<tr>
<td>Kashyap, Nitu</td>
<td>Kim, Min</td>
<td>Koss, Michele</td>
</tr>
<tr>
<td>Kashyap, Rahul</td>
<td>Kim, Min</td>
<td>Kotov, Alexander</td>
</tr>
<tr>
<td>Kasthurirathne, Suranga Nath</td>
<td>Kim, George</td>
<td>Kouril, Michal</td>
</tr>
<tr>
<td>Kato-Lin, Yi-Chin</td>
<td>Kimbrough, John</td>
<td>Krall, Michael</td>
</tr>
<tr>
<td>Kaufman, David</td>
<td>Kimura, Eizen</td>
<td>Krauss, John</td>
</tr>
<tr>
<td>Kaushik, Sashank</td>
<td>King, Brent</td>
<td>Kreuzthaler, Markus</td>
</tr>
<tr>
<td>Kavuluru, Ramakanth</td>
<td>Kirbiyik, Uzay</td>
<td>Krishnan, Rahul G</td>
</tr>
<tr>
<td>Keller, Dan</td>
<td>Kirkendall, Eric</td>
<td>Krive, Jacob</td>
</tr>
<tr>
<td>Kelley, Marjorie M</td>
<td>Kite, Bobbie</td>
<td>Kroth, Philip J</td>
</tr>
<tr>
<td>Kendall, Logan</td>
<td>Klann, Jeffrey Gordon</td>
<td>Krousel-Wood, Marie</td>
</tr>
<tr>
<td>Ketchum, Andrea</td>
<td>Kleinberg, Samantha</td>
<td>Kuhn, Thomson M</td>
</tr>
<tr>
<td>Khalifa, Aly MA</td>
<td>Kleyner, Yelena</td>
<td>Kum, Hye-Chung</td>
</tr>
<tr>
<td>Khan, Naqi Ali</td>
<td>Klinger, Elissa V</td>
<td>Kumar, Manish</td>
</tr>
<tr>
<td>Khan, Fauzia</td>
<td>Knoop, Sarah E</td>
<td>Kunisch, Joseph M</td>
</tr>
<tr>
<td>Khare, Ritu</td>
<td>Ko, Kelly J.</td>
<td>Kunjan, Kislaya</td>
</tr>
<tr>
<td>Kharrazi, Hadi</td>
<td>Koch, Sabine</td>
<td>Kuo, Alex</td>
</tr>
<tr>
<td></td>
<td>Kohler, Julianna</td>
<td>Kurc, Tahsin</td>
</tr>
</tbody>
</table>
AMIA 2015 REVIEWERS

Kuroda, Tomohiro
Kury MD, Fabricio Sampaio Peres
Kushniruk, Andre William
Kuusisto, Finn
Kuziemsky, Craig
Kwon, Bum Chul
Labkoff, Steven E
Lai, Feipei
Lai, Kenneth
Lake, Erica
Lalley, Catherine R
Lam, Cesar
Landais, Paul
Landis-Lewis, Zach
Lantz, Eric
LaSalle, Bernie
Lau, Lee Min
Lauritzen, Thomas
LaVenture, Marty
Lazar, Amanda
Leaman, Robert
Ledieu, Thibault
Lee, Ying-Li
Lee, Young Ji
Lee, Jinha
Lee, Mikyoung
Lee, Jaehoon
Lee, Kathy
Lee, Eva K
Lee, Yuyung
Lee, Victor C
Lee, JaeHo
Lee, Joon
Lefever, Els
Leftwich, Russell
Lele, Omkar
Lemke, Klaus
Lenert, Leslie Andrew
Leroy, Gondy
Leung, Tiffany
Leviss, Jonathan
Levy, Bruce
Li, Ying
Li, Man
Li, Qi
Li, Haomin
Li, Peter W
Li, Muqun
Li, Haiquan
Li, Dingcheng
Liang, Xi
Liaw, Siaw-Teng
Lichtner, Valentina
Liebeskind, David S
Likumahuwa, Sonja
Lim Choi Keung, Sarah N.
Lin, Jau-Huei
Lin, Jia-Ling
Lin, King-Ip
Lin, Yanhua
Lin, Yu
Lin, Chen
Line, Colleen
Lingren, Todd
Liu, Ruijue
Liu, Yang
Liu, Feifan
Liu, Mei
Liu, Jie
Liu, Yan
Liu, Xiao
Liu, Wing
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu, Haifeng</td>
<td>MacLeod, Haley</td>
<td>Markey, Mia K</td>
</tr>
<tr>
<td>Liu, Yuzhe</td>
<td>MacPherson, Jamie Ian</td>
<td>Marquard, Jenna</td>
</tr>
<tr>
<td>Lober, William B</td>
<td>Magee, Daniel J</td>
<td>Marshall, M. Scott</td>
</tr>
<tr>
<td>Lockett-Brown, Linda</td>
<td>Magrabi, Farah</td>
<td>Marshall, Robert Carter</td>
</tr>
<tr>
<td>Loehnhardt, Benjamin</td>
<td>Mahnke, Andrea</td>
<td>Marshall MS RN, Jean B</td>
</tr>
<tr>
<td>Long, Jie</td>
<td>Mahouli, Malika</td>
<td>Marsolo, Keith</td>
</tr>
<tr>
<td>Loos Dr., Stefan</td>
<td>Malin, Bradley</td>
<td>Martin, Danny</td>
</tr>
<tr>
<td>Lopetegui, Marcelo A</td>
<td>Mallow, Jennifer A</td>
<td>Martinez, Karen</td>
</tr>
<tr>
<td>Lopez Campos, Guillermo Hugo</td>
<td>Maloney, Christopher G</td>
<td>Martinez Costa, Catalina</td>
</tr>
<tr>
<td>López-García, Pablo</td>
<td>Malovini, Alberto</td>
<td>Martinez-Castellanos, Maria-Ana</td>
</tr>
<tr>
<td>Lorenzi, Virginia</td>
<td>Manaktala, Sharad</td>
<td>Massoudi, Barbara L</td>
</tr>
<tr>
<td>Lou, Jennie Q</td>
<td>Manard, William Troy</td>
<td>Matsumura, Yasushi</td>
</tr>
<tr>
<td>Loubser, Paul Gerhard</td>
<td>Mancini, Felipe</td>
<td>Mattmann, Christian A</td>
</tr>
<tr>
<td>Lourie, Eli M.</td>
<td>Mandel, Joshua C</td>
<td>May, John</td>
</tr>
<tr>
<td>Lu, Zhiyong</td>
<td>Mani, Subramani</td>
<td>McBride, Simon Jeffrey</td>
</tr>
<tr>
<td>Lucas, Joseph</td>
<td>Manion, Frank J</td>
<td>McCallie Jr, David</td>
</tr>
<tr>
<td>Lucena, Brian</td>
<td>Manning, John D</td>
<td>McCarthy, Lucas</td>
</tr>
<tr>
<td>Lucero, Robert James</td>
<td>Manos, Eva LaVerne</td>
<td>McClay, James C</td>
</tr>
<tr>
<td>Lulejian, Armine</td>
<td>Mao, Yuqing</td>
<td>McClure, Robert C</td>
</tr>
<tr>
<td>Luo, Jake</td>
<td>Maram, Naveen</td>
<td>McCormack, James L</td>
</tr>
<tr>
<td>Luyckx, Kim</td>
<td>Marceglia, Sara Renata</td>
<td>McCourt, Brian J.</td>
</tr>
<tr>
<td>Lyalin, David</td>
<td>Marcial, Laura Haak</td>
<td>McCoy, Allison Beck</td>
</tr>
<tr>
<td>Lyles, Courtney</td>
<td>Marin FACMI, Heimar de Fatima</td>
<td>McCoy Jr, Thomas H</td>
</tr>
<tr>
<td>Lyon, Lawrence</td>
<td>Marini, Simone</td>
<td></td>
</tr>
<tr>
<td>AMIA 2015 REVIEWERS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McCrary, Laura</td>
<td>Meystre, Stephane</td>
<td>Mohan, Vishnu</td>
</tr>
<tr>
<td>McCullagh, Lauren</td>
<td>Michalkiewicz, Michael</td>
<td>Mohan, Santosh</td>
</tr>
<tr>
<td>McDaniel, Anna</td>
<td>Michalowski, Martin</td>
<td>Moisa, Cristina</td>
</tr>
<tr>
<td>McDonald, Clement J</td>
<td>Michalowski, Wojtek</td>
<td>Mon, Donald T.</td>
</tr>
<tr>
<td>McDuffie, Michael Tommie</td>
<td>Michel, Jeremy</td>
<td>Monroe, Megan</td>
</tr>
<tr>
<td>McGarvey, Sunanda</td>
<td>Mikles, Sean</td>
<td>Monsen, Karen A.</td>
</tr>
<tr>
<td>McGinn, Roseanna J</td>
<td>Miksch, Tim</td>
<td>Moore, Susan L.</td>
</tr>
<tr>
<td>McGrath, Scott Patrick</td>
<td>Miller, Timothy A</td>
<td>Moosavinasab, Soheil</td>
</tr>
<tr>
<td>McManus, M. Sue</td>
<td>Miller, Christopher M</td>
<td>Mora-Jiménez, Inmaculada</td>
</tr>
<tr>
<td>McNamara, Mary</td>
<td>Miller, Andrew D</td>
<td>Moreno, Ramon Alfredo</td>
</tr>
<tr>
<td>Medlock, Stephanie</td>
<td>Millery, Mari</td>
<td>Moreno Franco, Pablo</td>
</tr>
<tr>
<td>Meeker, Daniella</td>
<td>Min, Hua</td>
<td>Morey, Angela</td>
</tr>
<tr>
<td>Mehrabi, Saeed</td>
<td>Miñarro-Gimenez, Jose Antonio</td>
<td>Morgan, Stephen James</td>
</tr>
<tr>
<td>Mehta, Prachi</td>
<td>Miotto, Riccardo</td>
<td>Mork, Peter</td>
</tr>
<tr>
<td>Mei, Jing</td>
<td>Miraglia, Simone</td>
<td>Mork, James G</td>
</tr>
<tr>
<td>Meigs, Michelle</td>
<td>Mirkovic, Jelena</td>
<td>Morland, Andrew J</td>
</tr>
<tr>
<td>Melton, Brittany</td>
<td>Mishra, Sonali R</td>
<td>Morlane-Hondere, François</td>
</tr>
<tr>
<td>Mendes, David José Murteira</td>
<td>Misra, Satish</td>
<td>Morley, Katherine</td>
</tr>
<tr>
<td>Mendonca, Eneida A</td>
<td>Mittu, Ranjeev</td>
<td>Morris, Jonathan A</td>
</tr>
<tr>
<td>Meng, Frank</td>
<td>Mo, Huan</td>
<td>Mosa, Abu Saleh Mohammad</td>
</tr>
<tr>
<td>Mentis, Helena M</td>
<td>Mobarek, Dalia A</td>
<td>Moss, Laura</td>
</tr>
<tr>
<td>Merrill, Jacqueline</td>
<td>Moen, Anne</td>
<td>Mossman, Ned</td>
</tr>
<tr>
<td>Metts, Christopher Lloyd</td>
<td>Mohammad, Amir</td>
<td>Mowery, Danielle</td>
</tr>
<tr>
<td>Metzger, Marie Helene</td>
<td></td>
<td>Mukherjee, Sukrit</td>
</tr>
</tbody>
</table>
AMIA 2015 REVIEWERS

Multak, Nina Loren  Nelson, Stuart James  Olinger, Lois
Mumba, Soyapi L  Neveol, Aurelie  Oliveira, Marilia Eutimia de Souza
Muñoz, Emir  Newbold, Susan K  Oliveira, Sergio Pacheco de
Murphy, Alison R  Ng, Kenney  Ologeanu Taddei, Roxana
Myerburg, Stuart  Nguyen, Anthony  Ong, Irene
Myers, Risa  Nguyen, Thanh Minh  Ong, Mei-Sing
Myneni, Sahiti  Nguyen, Vickie  Oniki, Thomas Alan
Nachimuthu, Senthil K  Nguyen, Lien B.  Oliveira, Sergio Pacheco de
Nadkarni, Girish Nitin  Nguyen, Phung Anh  Orreggio, Giordi
Naeymi-Rad, Frank  Nho, Kwangsik T  Overby, Casey Lynnette
Nagle, Lynn M  Nikolova, Ivelina Mircheva  Overhage, J. Marc
Nahm, Eun-Shim  Norcio, Anthony  Oyekola, Latifat Adebanke
Nair, Bala G  Nosal, Sarah C  Ozeran, Larry
Nakashima, Naoki  Novak, Laurie Lovett  Ozery-Flato, Michal
Nakayama, Masaharu  Nyström, Mikael  Ozbeknak, Mustafa
Nambisan, Priya  Nyulas, Csongor  Pacheco, Jennifer Allen
Nandigam, Hari Krishna  Obara, Piotr  Page, Jeni
Nanji, Karen C  Obradovic, Zoran  Page, David
Narou, Scott P  O'Bryan, Kevin  Painter, Ian
Naveed, Muhammad  Ochs, Christopher  Pakhomov, Serguei VS
Nayak, Manan  Odebode, Iyanuoluwa Emmanuel  Palchuk, Matvey B
Nazi, Kim Marie  Odlum, Michelle  Pan, Eric Chungren
Nelson, Scott D  Okhmatovskaia, Anya  Pandita, Deepti
Nelson, Greg  Okun, Sally  Panteleyeva, Natalya B
AMIA 2015 REVIEWERS

Papagari Sangareddy, Sridhar R  Pelletier, Alexandra  Prakash, Gouri
Park, Anne  Pendergrass, Sarah A  Pressler, Taylor R
Park, Albert  Pentakalos, Odysseas
Park, Jung In  Pepper, Catherine
Parkulo, Mark A  Perri, Seneca I
Parry, Dave  Perrin, Caroline
Partridge, Brittany  Peters, Steve G
Pasupathy, Kalyan  Petersen, Carolyn
Patel, Aman N  Peterson, Josh F
Patel, Jharana  Petrova, Mila
Patel, Vishal N  Petzel, Sue v
Paterno, Marilyn D  Pfeffer, Michael Adam
Paton, Chris  Phansalkar, Shobha
Patrick, Timothy B  Phillips, Andrew
Pattanayak, Arka  Phillips, Mark H
Patterson, Olga V  Pina, Jamie
Payne, Philip  Piscotty Jr., Ronald John
Peabody, Andrea L  Pivovarov, Rimma
Pedersen, Ted  Platt, Jodyn
Peek, Niels  Poikonen, John
Peeples, Malinda  Poker, Anna
Peissig, Peggy L.  Pollack, Ari
Pell, Jonathan  Popescu, Mihail
Poterack, Karl  Ranade Kharkar, Pallavi
Randell, Rebecca
<table>
<thead>
<tr>
<th>AMIA 2015 REVIEWERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rashidee, Ali</td>
</tr>
<tr>
<td>Rasmussen, Luke</td>
</tr>
<tr>
<td>Ravkic, Irma</td>
</tr>
<tr>
<td>Ravvaz, Kouros</td>
</tr>
<tr>
<td>Ray, Bisakha</td>
</tr>
<tr>
<td>Razavian, Narges</td>
</tr>
<tr>
<td>Rea, Susan</td>
</tr>
<tr>
<td>Rector, Alan L</td>
</tr>
<tr>
<td>Redd, Travis Kenneth</td>
</tr>
<tr>
<td>Redd, Doug</td>
</tr>
<tr>
<td>Reddy, Madhu</td>
</tr>
<tr>
<td>Redley, Bernice</td>
</tr>
<tr>
<td>Reed, Mary</td>
</tr>
<tr>
<td>Reeder, Blaine</td>
</tr>
<tr>
<td>Reeves, Ruth M</td>
</tr>
<tr>
<td>Resetar, Ervina</td>
</tr>
<tr>
<td>Rezaie, Mohammadreza</td>
</tr>
<tr>
<td>Richards, Janise</td>
</tr>
<tr>
<td>Rigla, Mercedes</td>
</tr>
<tr>
<td>Riley, William T</td>
</tr>
<tr>
<td>Rindflesch, Thomas</td>
</tr>
<tr>
<td>Rios Piedra, Edgar A.</td>
</tr>
<tr>
<td>Rizvi, Rubina Fatima</td>
</tr>
<tr>
<td>Robbins, Jeffrey</td>
</tr>
<tr>
<td>Roberts, Angus</td>
</tr>
<tr>
<td>Roberts, Kirk</td>
</tr>
<tr>
<td>Rocca, Mitra</td>
</tr>
<tr>
<td>Roche, Erin</td>
</tr>
<tr>
<td>Rochefort, Christian M.</td>
</tr>
<tr>
<td>Roden, Robert</td>
</tr>
<tr>
<td>Roderer, Nancy K</td>
</tr>
<tr>
<td>Roderick, Oleg E</td>
</tr>
<tr>
<td>Rodríguez González, David</td>
</tr>
<tr>
<td>Rogers, Michelle L</td>
</tr>
<tr>
<td>Rogers, Michelle L</td>
</tr>
<tr>
<td>Rojo-Álvarez, José Luis</td>
</tr>
<tr>
<td>Romero, Mary Rachel</td>
</tr>
<tr>
<td>Rosenau, Paul T</td>
</tr>
<tr>
<td>Rosenman, Marc</td>
</tr>
<tr>
<td>Rosen-Zvi, Michal</td>
</tr>
<tr>
<td>Ross, Mindy K</td>
</tr>
<tr>
<td>Rothman, Michael</td>
</tr>
<tr>
<td>Rothman, Brian S.</td>
</tr>
<tr>
<td>Roychoudhury, Shoumik</td>
</tr>
<tr>
<td>Rubinos, Mary Ann T</td>
</tr>
<tr>
<td>Rudin, Robert S</td>
</tr>
<tr>
<td>Rule, Adam</td>
</tr>
<tr>
<td>Runger, George</td>
</tr>
<tr>
<td>Rusincovitch, Shelley A</td>
</tr>
<tr>
<td>Russ, Alissa L</td>
</tr>
<tr>
<td>Saba, Virginia K.</td>
</tr>
<tr>
<td>Sabbatini, Renato Marcos Endrizzi</td>
</tr>
<tr>
<td>Safran, Charles</td>
</tr>
<tr>
<td>Sairamesh, Jakka</td>
</tr>
<tr>
<td>Saiyed, Salim</td>
</tr>
<tr>
<td>Sakaguchi, Farrant</td>
</tr>
<tr>
<td>Salimi, Mandana</td>
</tr>
<tr>
<td>Salimian, Parissa K</td>
</tr>
<tr>
<td>Sands, Daniel Zev</td>
</tr>
<tr>
<td>Sarantis, Demetrios</td>
</tr>
<tr>
<td>Sarcevic, Aleksandra</td>
</tr>
<tr>
<td>Sarkar, Indra Neil</td>
</tr>
<tr>
<td>Sarker, Abeed</td>
</tr>
<tr>
<td>Sarma, Karthik</td>
</tr>
<tr>
<td>Venkataraman</td>
</tr>
<tr>
<td>Sarmiento, Raymond Francis</td>
</tr>
<tr>
<td>Schaeffer, Molly</td>
</tr>
<tr>
<td>Schaffer, Jonathan Lawrence</td>
</tr>
<tr>
<td>Schaub, Marc Andreas</td>
</tr>
<tr>
<td>Schendel, Eric</td>
</tr>
<tr>
<td>Scheufele, Elisabeth</td>
</tr>
<tr>
<td>Schlegel, Daniel R</td>
</tr>
</tbody>
</table>
AMIA 2015 REVIEWERS

Schmitt, Charles
Schnall, Rebecca
Schneider, Eric
Schreiber, Richard
Schubert, Carolyn
Schulz, Stefan
Schwartz, Jessica
Schwei, Kelsey M
Scotch, Matthew
Scott, Jean
Seckman, Charlotte A
Selva, Luis E
Seng, Michele Anando
Seo, Hwa Jeong
Sepulveda, Jorge
Seror, Ann
Sevenster, Merlijn
Sevilla Berrios, Ronaldo A
Shabah, Abdo
Shaban-Nejad, Arash
Shaffer, Jonathan
Shahar, Yuval
Shahrian Varnousfaderani, Ehsan
Shang, Ning
Shatkay, Hagit
Shea, Kimberly
Sheets, Lincoln
Sheide, Amy
Sheikh, Aziz
Sheikhtaheri, Abbas
Shen, Shuying
Shenson, Jared Andrew
Shenvi, Edna C
Shewbridge, Rita
Shivade, Chaitanya
Sideris, Konstantinos
Siff, Jonathan
Silva, Luis Manuel
Simms, Andrew
Simon, Gyorgy
Simpao, Allan F.
Sims, Maria
Singleton, Kyle W
Sittig, Dean F
Skapik, Julia
Skiba, Diane J
Skrøvseth, Stein Olav
Slater, Elyse
Smalheiser, Neil R.
Smith, Catherine Arnott
Sockolow, Paulina
Soguero-Ruiz, Cristina
Solomonides, Anthony Eugene
Somai, Melek
Soman, Sandeep
Song, Yiqiang
Sorace, James
Sottara, Davide
Soulakis, Nicholas
South, Brett Ray
Souza, Brunno Wagner Lemos de
Spallek, Heiko
Speier, William
Sperrin, Matthew
Spyropoulos, Basile P.
Srinivas, Preethi
Staes, Catherine J
Stein, Daniel
Steinbach, Michael
Stepanek, Richard
Stephan, Christina L
AMIA 2015 REVIEWERS

Stevenson, Mark
Stiglic, Gregor
Su, Wu-Chen
Suh, Edward H.
Sujansky, Walter
Sumner II, Walton
Sun, Mengmeng
Sutton, Joseph A
Sward, Katherine A
Swartz, Jordan
Syed, Zeeshan
Tabernik, Ty J
Takemoto, Steve
Takesue, Blaine
Tamang, Suzanne
Tan, Xing
Tandon, Sanjeev
Tannier, Xavier
Tant, Elizabeth M.
Tao, Donghua
Tao, Cui
Tenenbaum, Jessica D
Theera-Ampornpunt, Nawanan
Thirumalai, Vijayakumar
Thomas, Sarah Katie
Thorpe, Jane Hyatt
Thum III, Frederick Lorenz
Thyvalikakath, Thankam P.
Tiase, Victoria
Timm, John
Timperi, Ralph
Tokachichu, Priyaranjan
Tolentino, Herman
Topaz, Maxim
Torresani, Michele
Torres-Urquidy, Miguel Humberto
Toscos, Tammy
Travers, Jasmine
Tredinnick, Ross
Tremblay, Monica Chiarini
Tripp, Jacob S.
Trocki, Rebecca R
Tsarkov, Dmitry
Tsui, Rich
Tu, Samson
Turpin, Rebecca Little
Tweedy, Carolyn G.
Unertl, Kim M
Ursu, Oleg
Vail, Laura
Valdez, Rupa
Van Auken, Kimberly M
Van Mulligen, Erik M.
VanHouten, Jacob
Verspoor, Karin
Vest, Joshua R
Vizer, Lisa
Vollmer, Anne-Maria
Vorderstrasse, Allison
Vydiswaran, V.G.Vinod
Wagholikar, Amol
Waks, Zeev
Wald, Jonathan S
Walkerk, Janice
Wallace, Byron C
Wallace, Byron
wang, Xiaoming
Wang, Xiang
Wang, Dongwen
Wang, Shuang
Wang, Xiaoyan
<table>
<thead>
<tr>
<th>Wang, Liqin</th>
<th>Wright, Julie A</th>
<th>Yetisgen, Meliha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang, Fei</td>
<td>Wright, Melanie C</td>
<td>Yim, Wen-Wai</td>
</tr>
<tr>
<td>Wang, Fusheng</td>
<td>Wright, Adam</td>
<td>Yorks, Melissa</td>
</tr>
<tr>
<td>Wangia-Anderson, Victoria</td>
<td>Wu, Han</td>
<td>Yoshida, Eileen</td>
</tr>
<tr>
<td>Ward, Merry</td>
<td>Wu, Cai</td>
<td>Yu, Zhiguo</td>
</tr>
<tr>
<td>Warner, Jeremy L.</td>
<td>Wu, Yirong</td>
<td>Yu, Fei</td>
</tr>
<tr>
<td>Warner, Phillip B.</td>
<td>Wu, Danny TY</td>
<td>Yu, Yiqin</td>
</tr>
<tr>
<td>Webber, Emily C.</td>
<td>Wyatt, Matthew</td>
<td>Yu, Kun-Hsing</td>
</tr>
<tr>
<td>Weber, Jens</td>
<td>Xia, Yuni</td>
<td>Yu Jr., Feliciano Buenviaje</td>
</tr>
<tr>
<td>Weber, Patrick</td>
<td>Xie, Yue</td>
<td>Yusuf, Rafeek A</td>
</tr>
<tr>
<td>Wei, Chih-Hsuan</td>
<td>Xu, Songhua</td>
<td>Zachary, Iris</td>
</tr>
<tr>
<td>Wei, Wei-Qi</td>
<td>Xu, Yaomin</td>
<td>Zai, Adrian</td>
</tr>
<tr>
<td>Weir, Charlene</td>
<td>Xu, Hua</td>
<td>Zeng, Jia</td>
</tr>
<tr>
<td>Welch, Brandon M</td>
<td>Yackel, Thomas R</td>
<td>Zhang, Shaodian</td>
</tr>
<tr>
<td>Westra, Bonnie L</td>
<td>Yan, Su</td>
<td>Zhang, Yongkang</td>
</tr>
<tr>
<td>Whetzel, Patricia L</td>
<td>Yang, Shiming</td>
<td>Zhang, Jing</td>
</tr>
<tr>
<td>Whiting, Kandace Mikel</td>
<td>Yang, Haodong</td>
<td>Zhang, Songmao</td>
</tr>
<tr>
<td>Wiens, Jenna</td>
<td>Yang, Yang</td>
<td>Zhang, Rui</td>
</tr>
<tr>
<td>Wilcox, Adam</td>
<td>Yanover, Chen</td>
<td>Zhang, Mingyuan</td>
</tr>
<tr>
<td>Wiley, Laura K</td>
<td>Yao, Lixia</td>
<td>Zhao, Jing</td>
</tr>
<tr>
<td>Wiley, Matthew T</td>
<td>Ye, Nan</td>
<td>Zhe, Tian</td>
</tr>
<tr>
<td>Wilk, Szymon</td>
<td>Ye, Ye</td>
<td>Zheng, Ling</td>
</tr>
<tr>
<td>Winden, Tamara</td>
<td>Ye, Zhan (Harold)</td>
<td>Zheng, Jiaping</td>
</tr>
<tr>
<td>Wood, Grant M</td>
<td>Yen, Po-Yin</td>
<td>Zhou, Xiaomu</td>
</tr>
</tbody>
</table>
Zhou, Li

Zimolzak, Andrew J

Zulman, Donna M.

Zweigenbaum, Pierre
Notice

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

David Albers¹, PhD, Rimma Pivovarov¹, MA, J Michael Schmidt², PhD,
Noémie Elhadad¹, PhD, George Hripcsak¹, MD MS
¹Department Biomedical Informatics, Columbia University, New York, NY
²Department of Neurology, Columbia University, New York, NY

Introduction
A core task when developing data-driven tools on electronic health record (EHR) data is to model the raw observations in a way that is computationally feasible and that represents the underlying phenomena that generated the observations [1]. While non-parametric models do represent observations in an accurate fashion, their parametric alternative models have the advantages of interpretability (e.g., mean, shape, variance, are all parameters which help describe or summarize raw observations) and known, established properties that can be leveraged (e.g., conjugacy of different known distribution families). Here we study the question of parametric model selection, and focus on laboratory test results (i.e., a range of laboratory tests and the many observed values for these tests in the EHR). Prior work has observed that the health care process (HCP) [1], as defined through measurement context [9] and measurement patterns [2], can influence how EHR data are represented as statistical parameterizations. We present a methodology for model selection of EHR laboratory tests. To assess the value of our methodology, we experiment on two different datasets, generated through different clinical healthcare processes (primary care vs. intensive care). Furthermore, for laboratory tests where the underlying physiology is well established, we examine the validity of the selected model to preserve their underlying physiological properties.

Methods
Study was carried over two cohorts from different contexts. The first includes EHR data collected during a stay in a neurological intensive care unit (ICU) from patients who are comatose. The second cohort (AIM) comprises longitudinal records of patients who visit regularly an outpatient clinic, as well as get admitted and visit the emergency department. The HCPs underlying these data are very different: in the ICU cohort, data are randomly sampled at regular intervals largely independent of the overall health state of patients, while in the AIM cohort, data are sampled primarily during visits, distributed through time as patients gets sick.

We use a kernel density estimate (KDE) [2] of patient laboratory data as a non-parametric gold standard model. We then fit the data to 10 different established, parameterized models (e.g., Gaussian, Gamma [4], Generalized Extreme Value (GEV) [5], etc.). The most representative model is selected using a variation of an information criterion based goodness-of-fit measure, the KL-divergence [3] between the gold standard KDE and our parametric models. This model evaluation technique can be used in a high throughput setting and is applicable to a range of applications from phenotyping to cohort selection. Finally, we verify that the model evaluation methodology (KL-divergence) selects a realistic model by showing that for glucose the best parametric model preserved known physiologic relationships, specifically, that mean-like and variance-like quantities increase simultaneously [7-8].

Results
For the ICU cohort, the parameterization that minimized the KL-divergence was the gamma distribution. The mean-like and variance-like gamma parameters increased simultaneously (Fig. 1). In contrast, no other model, including the next best model (Fig. 2) reproduced that physiologic relationship. That the gamma distribution was chosen is reassuring because it is the maximum entropy model for any system whose mean-like and variance-like quantities are linearly related [4]. In contrast, for the AIM cohort, the GEV was the parameterization that minimized the KL-divergence. The mean-like and variance-like GEV parameters increase simultaneously (Fig. 3), reproducing the known physiologic feature. No other model reproduced that physiologic relationship without human intervention (Fig. 4). The GEV is the generating process for the measurement of extremes of a distribution [5], implying that the AIM data are generated by a process that measures the extremes of the physiologic process, revealing effects of the HCP.

Discussion
We have shown that in different clinical contexts, different models of data may be appropriate, and we believe that the cause is not so much a change in physiology, but a change in the way health is measured in the different contexts. That is, the HCP causes different but measurable biases in the data. In the ICU, the HCP measurement function does not appear to have a strong influence on glucose measurements. This implies that we can treat and use ICU data more like continuously sampled physiologic data. In contrast, the process that generates or collects data from outpatient, or mixed outpatient/inpatient settings, does influence the data collected. If the correct model is not chosen, the physiologically expected relationships between parameters are lost. Fortunately, selecting a model that minimizes the KL-divergence between a model and the KDE, a methodology that can be used in a high throughput setting, can correct for the bias due to context. Meaning, selecting the model that minimizes the KL-divergence seems to minimize the impact of the HCP on the modeling of the data, allowing us to identify the most physiologically faithful representation of the physiologic process within EHR data.

Acknowledgements We acknowledge NLM grant R01 LM06910, NSF award 1344668, and NLM T15 LM007079.
References

Analysis and Classification of Patient Safety Reports in Computerized Prescriber Order Entry (CPOE) Systems and Refinement of a New Taxonomy for Classification of CPOE-Related Medication Errors

Mary G. Amato, PharmD, MPH 1,2, Alejandra Salazar, PharmD 1, Thu-Trang Hickman, MPH 1, Arbor J.L. Quist 1, Lynn Volk, MHS 3, Adam Wright, PhD 1,4, Dustin McEvoy 3, Sarah P. Slight, PhD, PGDip 1,5, David W. Bates, MD, MSc 1,4, Gordon D. Schiff, MD 1,4
1 Brigham and Women’s Hospital, Division of General Internal Medicine and Primary Care, Boston, MA, USA; 2 MCPHS University, Boston, MA, USA; 3 Partners HealthCare, Information Systems, Wellesley, MA, USA; 4 Harvard School of Medicine, Boston, MA, USA; 5 School of Medicine, Pharmacy and Health, The University of Durham, UK; 6 Harvard School of Public Health, Boston, MA, USA

Introduction

CPOE has been shown to reduce medication error rates overall, but can also facilitate the risk of some types of prescribing errors.1 Obtaining the full safety benefits of CPOE demands that organizations develop approaches to identify and address the causes of these new errors. The aim of this study was to classify and analyze errors related to CPOE with ongoing development and refinement of a taxonomy classification system for CPOE-related errors.

Methods

We reviewed medication-related patient safety and error reports from January-December 2013 (n = 2522) from six sites that participated in an FDA-sponsored project examining safety issues in CPOE systems. Two pharmacists reviewed the error narrative reports and identified those that occurred in the ordering/prescribing phase of medication use that were related to the use of CPOE. For those related to CPOE, we assessed whether CPOE facilitated (i.e. actively contributed to) the error or failed to prevent the error (i.e. CPOE system improvements may have prevented the error). A taxonomy which classified errors related to CPOE developed in a previous project was revised and updated in order to classify the reports regarding what happened to the patient, what happened in CPOE, and, if it could be determined from the report, why the error occurred.2 The taxonomy was iteratively refined as more cases were reviewed and different types of errors were found.

Results

Of the 2522 medication error reports reviewed, 1214 (48.1%) were deemed not related to CPOE mainly because they did not occur in the ordering phase or did not involve an electronic order. Of the remaining 1308 reports, CPOE was thought to facilitate the error in 171 (13.1%) and failed to prevent the error in 1137 (86.9%). The final version of the taxonomy included 14 categories describing what happened to the patient, 40 categories describing what went wrong when entering the order in CPOE, and 71 potential factors contributing to why the error occurred. The three most frequently cited reasons for what happened when entering the order in CPOE were “Other: Entered order not routed to/received at intended destination”, “Dose: Ordered wrong dose or strength”, and “Drug: Duplicate order: Same exact drug” (Table 1). A wide variation was seen in the format, categorization, and quality of the narratives, making it difficult to use the taxonomy to classify what happened and why in some cases.

Discussion

The leading errors related to CPOE were those related to transmission errors, erroneous dose, and duplicate orders. Our review of error report narratives was useful in further refining a taxonomy classifying CPOE-related errors. A more standardized safety reporting process using a common taxonomy in systems across healthcare facilities would enable healthcare systems to learn more about the causes of and implement prevention strategies for these errors.
Table 1. Top 15 “What Happened in CPOE?” Categories of CPOE-Related Errors Across All Six Sites, 2013

<table>
<thead>
<tr>
<th>What Happened in CPOE?</th>
<th>CPOE Facilitated (n = 171)</th>
<th>CPOE Failed to Prevent (n = 1137)</th>
<th>Total (%) (n =1308)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other: Entered order not routed to/received at intended destination</td>
<td>35</td>
<td>405</td>
<td>440 (33.6)</td>
</tr>
<tr>
<td>Dose: Ordered wrong dose or strength</td>
<td>8</td>
<td>149</td>
<td>157 (12.1)</td>
</tr>
<tr>
<td>Drug: Duplicate order: Same exact drug</td>
<td>15</td>
<td>125</td>
<td>140 (10.7)</td>
</tr>
<tr>
<td>Drug: Ordered drug that was inappropriate or contraindicated (by lab, disease, age, pregnancy, interactions with other drug, or patient’s explicit refusal of drug)</td>
<td>3</td>
<td>48</td>
<td>51 (4.0)</td>
</tr>
<tr>
<td>Drug: Ordered wrong dosage form (IR, ER, SR, XR; tablets, capsules; oral, topical) or formulation</td>
<td>2</td>
<td>43</td>
<td>45 (3.4)</td>
</tr>
<tr>
<td>Drug: Ordered wrong drug</td>
<td>2</td>
<td>42</td>
<td>44 (3.4)</td>
</tr>
<tr>
<td>Drug: Failure to order drug: Failure to renew or reorder drug (including home, chronic, held medications, antibiotic renewals)</td>
<td>0</td>
<td>38</td>
<td>38 (2.9)</td>
</tr>
<tr>
<td>Time: Wrong time, schedule, or frequency entered</td>
<td>8</td>
<td>22</td>
<td>30 (2.3)</td>
</tr>
<tr>
<td>Drug: Ordered drug to which patient was allergic</td>
<td>0</td>
<td>29</td>
<td>29 (2.2)</td>
</tr>
<tr>
<td>Drug: Failure to order drug</td>
<td>0</td>
<td>27</td>
<td>27 (2.1)</td>
</tr>
<tr>
<td>Dose: Medication number or quantity problem: wrong number or quantity indicated on order</td>
<td>0</td>
<td>22</td>
<td>22 (1.7)</td>
</tr>
<tr>
<td>Drug: Failure to order drug: Failure to order an indicated drug or corollary order</td>
<td>1</td>
<td>18</td>
<td>19 (1.5)</td>
</tr>
<tr>
<td>Drug: Ordered wrong drug: Ordered LASA</td>
<td>1</td>
<td>14</td>
<td>15 (1.1)</td>
</tr>
<tr>
<td>Time: Wrong time, schedule, or frequency entered: wrong order date</td>
<td>2</td>
<td>13</td>
<td>15 (1.1)</td>
</tr>
<tr>
<td>Dose: Failure to order dose change</td>
<td>0</td>
<td>13</td>
<td>13 (1.0)</td>
</tr>
</tbody>
</table>

References

2. Schiff GD, Amato MG, Eguale T et al.. Computerized physician order entry-related medication errors: Analysis of reported errors and vulnerability testing of current systems. BMJ Qual Saf Published online first (16 Jan 2015) doi:10.1136/bmjqs-2014-003555.

This project is supported by the U.S. FDA (Contract: HHSF22320100008I, Task Order: HHSF22301005T).
A Review and Analysis of Rounding and Handoff Document Content in Inpatient Resident Physician Teams

Elliot Arsoniadis, MD\textsuperscript{1,2}, Rohini Khatri, MD\textsuperscript{2}, Jenna Marquard, PhD\textsuperscript{3}, Courtney Moors\textsuperscript{3}, Michael Kim, MD\textsuperscript{3,4}, Genevieve B. Melton, MD, PhD\textsuperscript{1,2}

\textsuperscript{1}Institute for Health Informatics, \textsuperscript{2}Department of Surgery, \textsuperscript{3}Medical School, and \textsuperscript{4}Department of Medicine, University of Minnesota, Minneapolis, MN

\textsuperscript{5}College of Engineering, University of Massachusetts, Amherst, MA

Introduction

In the seminal paper in 2004, van Eaton and colleagues described the content needs of residents regarding electronic rounding and handoff documents (1). Several common themes emerged from this work and subsequent studies (2,3). Despite the decade of literature that has followed since this work, standardization of content of electronic rounding or handoff documents to better inform EHR vendors is lacking. To formulate a basis for electronic rounding and handoff document standards and determine the current needs, we engaged in a three-part investigation: 1) searching the literature to determine content in EHR-based rounding and handoff documents, 2) examining the content of these documents at our own institution, and 3) interviewing resident physicians regarding their attitudes and opinions about present document content.

Methods

Review of the literature revealed at least one existing standard and several common themes in standard EHR rounding and handoff document content (4,5). To gain an understanding of the data elements in rounding and handoff documents in use at our own institution, we gathered 18 documents from 18 different clinical service teams and extracted content headings. We then normalized these to the Clinical Document Architecture (CDA) Continuity of Care Document (CCD) standard and other common data field themes obtained from the literature and tabulated the results. To gain insight into content evaluation by resident physicians, we engaged residents from Internal Medicine, Pediatrics, and Surgery (n=16) in semi-structured interviews. Open-ended questions regarding rounding and handoff document content value and perceived accuracy, as well as missing or desired elements, were asked and answers recorded. Sample rounding and handoff documents were utilized during the interviews. Interview content was analyzed and answers quantified.

Results

Only 3 out of 18 documents described \textit{Code Status} despite this information being highly reported of value in the literature. In contrast, 16 of 18 documents contained some form of documentation on \textit{Plan of Care}, consistent with the high value placed on anticipatory guidance in the literature. Surgical residents demonstrated greater consistency in utilization of data fields, with 4 of 11 categories utilized by 100\% of interviewed subjects. In contrast, neither Pediatric nor Internal Medicine team members unanimously employed any given data field. Internal Medicine teams also had the highest level of mistrust of data elements, with a near equal mix of manually entered and EHR populated elements. When asked about desired improvements, Internal Medicine and Surgery residents both dedicated more than 50\% of their comments to \textit{Medications} and \textit{Hospital Course}, which are fields that contain a near even mix of manually entered and EHR generated data. The predominant themes limiting user satisfaction in EHR generated documents related to timeliness of populated values compared to when the information is used, and excess of “noise” of data not perceived as useful, which precluded efficient identification of valued elements. In contrast, manually entered fields were nearly always considered suspect for accuracy of updates, and their success depended largely on daily work burden associated with maintaining their content.

Conclusion

There has yet to emerge a standard for EHR-based rounding and handoff documents. There appears to be significant overlap of resident physician needs with content elements published in the literature, as well as additional areas of development and improvement desired for these tools to optimize communication and use of these tools for patient care.
Table 1. Analysis of Existing Rounding and Handoff Documents

<table>
<thead>
<tr>
<th></th>
<th>Analysis of Existing Rounding and Handoff Documents from 18 Clinical Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Data Fields Extracted</td>
<td>72</td>
</tr>
<tr>
<td>Number of Normalized Fields</td>
<td>17</td>
</tr>
<tr>
<td>Analysis based on Document Construction</td>
<td>Word Document</td>
</tr>
<tr>
<td>Total Rounding and Handoff Documents</td>
<td>6</td>
</tr>
<tr>
<td>Most data fields in single document</td>
<td>22</td>
</tr>
<tr>
<td>Range of Normalized Fields</td>
<td>2-12</td>
</tr>
<tr>
<td>Median number of Normalized Fields</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Table 2. Analysis of Resident Report Content Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Pediatrics</th>
<th>Medicine</th>
<th>Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean utilization of fields in Rounding Tool</td>
<td>47%</td>
<td>44%</td>
<td>89% * (p value 0.045 compared with pediatrics; =0.0026 with Internal Medicine)</td>
</tr>
<tr>
<td>Mean utilization of fields in Handoff Tool</td>
<td>60%</td>
<td>41%</td>
<td>48%</td>
</tr>
<tr>
<td>Total Data elements not trusted (% EHR generated)</td>
<td>2 (0)</td>
<td>19 (58)</td>
<td>4 (0)</td>
</tr>
<tr>
<td>% of Comments dedicated to improving data fields related to medications and hospital course</td>
<td>0</td>
<td>57-75%</td>
<td>69-88%</td>
</tr>
</tbody>
</table>

Acknowledgements

This work was partially supported by National Science Foundation Award #CMMI-1150057 (JM) and the Agency for Healthcare Research and Quality Award #R01HS022085 (GM). The authors would also like to thank Fairview Health Services, University of Minnesota Physicians, and University of Minnesota Department of Surgery for support of this work.

References


STEMPOWERMENT: A Prototype Online Intervention to Improve Outcomes in Stem Cell Transplant Survivors

Martha Askins, PhD1, Jeffery McLaughlin2, Susan Peterson, PhD1, Dongsong Zhang, PhD3

1 MD Anderson Cancer Center, Houston, TX, 2 Radiant Creative Group, LLC, Houston, TX, 3 The University of Maryland Baltimore, County, Baltimore, Maryland.

Abstract

Stem cell transplantation survivors face numerous physical and psychosocial challenges during recovery. Low adherence to recovery behaviors can result in poor health outcomes and increased risk of death. We designed and evaluated a prototype online social intervention that utilized game mechanics to promote post-transplant self-management behaviors among AYA survivors; namely, medication adherence, hydration, and physical activity.

Background

Hematopoietic stem cell transplantation (SCT) is a life-threatening, intensive treatment regimen involving a lengthy hospital stay and a protracted recovery period. During recovery, SCT survivors encounter numerous physical and psychosocial barriers to adherence with medication regimens and maintenance of adequate hydration and physical activity levels; low adherence can result in serious consequences such as organ damage and/or chronic, severe graft vs. host disease. This is especially true for adolescent and young adult (AYA) SCT survivors, whose developing autonomy and enjoyment of peers is often juxtaposed with the constraints of cancer treatment. We evaluated a prototype online interactive intervention utilizing game mechanics to promote post-SCT self-management behaviors among AYA survivors; namely, medication adherence, hydration, and physical activity.

Methods

The prototype intervention implemented an online social network that incorporated game mechanics and collaborative problem solving. Technical merit, feasibility and usability of the intervention were evaluated using the framework of the Technology Acceptance Model (TAM), and incorporated constructs related to system performance, perceived enjoyment, behavioral intention to use the system, and usage behavior. Participants included AYA allogeneic SCT survivors who accessed the intervention site for up to 3 weeks; they subsequently completed an evaluation comprising a 52-item TAM measure and semi-structured interviews.

Results

16 AYA allogeneic SCT survivors participated; 52% were female, mean age was 22 years (range 18-28). The overall mean TAM score was 6.2 (SD=2.1, range 1-7); subscale scores ranged from 5.7 (lowest) on perceived usefulness to 6.9 (highest) on behavioral intention. Qualitative responses indicated that users valued the ability to create personalized profiles, learning self-management strategies, and social networking with AYA SCT survivors. Desired feature enhancements included more interactive activities and information on long-term coping after SCT.

Conclusions

Results showed high usability, feasibility and satisfaction ratings with the intervention, and a desire for long-term coping information post-SCT and more interactive features.
Assessing Variability in Breast Cancer Treatment Paths Using Frequent Sequence Mining

Ravi V. Atreya¹, Thomas A. Lasko, MD, PhD¹, Mia A. Levy, MD, PhD¹
¹Vanderbilt University School of Medicine, Nashville, TN

Introduction
Quality measurement of breast cancer treatment paths can help healthcare providers and organizations improve the delivery of care and engage with patients. A major component of tracking clinical quality is assessing the variability in care. In this work, we aim to use frequent sequence mining to identify frequently occurring event patterns to help us represent infrequently occurring treatment paths. It is important to identify infrequently occurring treatment paths to understand why deviations from the standard of care are occurring and what the resulting patient outcomes are. Given the proliferation of quality measures in breast cancer care, breast cancer centers are facing growing pressures to understand the treatment paths of their patients in order to improve the care they provide.

Methods
We selected adult women diagnosed with stage I-III breast cancer between 2000-2012 from the cancer registry at Vanderbilt. We extracted treatment event data from the cancer registry and focused on the four primary modalities of breast cancer treatment: surgery, chemotherapy, hormone therapy, and radiation therapy. We represented surgical events as either breast conserving surgeries or mastectomies. We generated treatment event sequences by placing events in sequential order based on their timestamps.

We used the SPADE event sequence mining method² in the arulesSequences R package³ to identify treatment event patterns within the event sequences. A treatment event pattern is a subset of an event sequence; for example, a “breast conserving surgery—radiation therapy” event pattern is a subset of a “breast conserving surgery—chemotherapy—radiation therapy” event sequence. We ranked the treatment event patterns by their frequency (0.5% cutoff) and calculated the number event sequences that the most frequent treatment event patterns were present in. This analysis is aimed at representing the variability in the treatment event sequences by using frequently occurring treatment event patterns to assess infrequently occurring treatment event sequences.

Results
We extracted 1528 patient records and their treatment events from the Vanderbilt cancer registry. We generated 265 unique treatment event sequences for the 1528 patients. Ninety percent of the patient population could be represented by the top 126 (47.5%) treatment event sequences. We plotted the distribution of treatment event sequences across our patient population (Figure 1, left). This demonstrates how a majority of the patient population shares a common set of treatment event sequences. The other 10% of the population, however, is represented by the bottom 139 (52.5%) treatment event sequences. This long tail in the distribution represents variability in breast cancer care.

We next identified 144 frequently occurring treatment event patterns using the SPADE sequence mining method and plotted their distribution across our patient population (Figure 1, right). Ninety percent of the patient population could be represented by the 52 (36.1%) most frequently occurring treatment event patterns of at least two events. Of the 153 patients (10% of population) represented by 139 treatment event sequences (the long tail), 95% were successfully represented by the seven (4.9%) most frequently occurring treatment event patterns of at least two events. The most frequent event pattern of at least two events (breast conserving surgery—radiation therapy) is an important quality measure in breast cancer care. Finally, we visually represented a subset of surgically relevant treatment event patterns in hierarchical fashion to show how frequently occurring event patterns can be joined together to generate longer, more infrequently occurring event patterns and, eventually, event sequences (Figure 2).

Discussion
We represented the variability in breast cancer care by assessing the distribution of treatment event sequences and aligning them with our frequently occurring treatment event patterns. Infrequent treatment event sequences (long tail) were represented by the seven most frequently occurring event patterns. The diverse set of treatment paths, could still be represented through a small set of treatment event patterns. We plan to cluster our treatment event sequences using the results of our frequent event sequence mining approach.
Figure 1. Distribution of treatment event sequences (left) and frequently occurring treatment event patterns (right). The left side shows 265 unique treatment event sequences ordered by frequency in the patient population. The right side shows 144 frequently occurring treatment event patterns ordered by frequency in the patient population. The color represents the first event in the sequence & pattern.

Figure 2. Hierarchy of frequently occurring treatment event patterns focused on surgical care. This shows how frequently occurring event pattern parents can be joined together to generate longer, more infrequently occurring event pattern children. While the BCS, Mast, C pattern only occurs in 3% of the population, its parents C,Mast & BCS,C occur in 12% & 21% of the population, respectively. Color scheme similar to Figure 1.

References
Six Important Requirements for Patient Hand-Off Application in Inpatient Hospital Setting

Soehi U. Ayubi, PhD, Alexandra C. Pelletier, MBA

1Innovation Acceleration Program, Boston Children’s Hospital, Boston, MA

Abstract
We present six important and required requirements for patient hand-off application in an inpatient setting using bring-your-own-device (BYOD). The requirements are learning lessons from developing and piloting a hand-off application at Boston Children’s Hospital.

Introduction
Caring for hospitalized patients with many team members in an inpatient environment is complex, and hand-off communications among the team members are a leading cause of medical errors.1 The members often have many tasks, divide their attention between multiple tasks and patients, and have high demands around communication and documentation.2 Physicians often use paper checklists to track their daily tasks and facilitate hand-offs.3 Paper is really convenient, but is a static, redundant, unstandardized, unshared, and insecure medium.4 Checklists are seldom available in commercial electronic medical records (EMRs) and often lack mobile, live-tracking components, and often provide only limited integration with EMRs.5 Non-EMR checklists applications are also available to emulate these paper task lists, but may not have required security or privacy controls, may not be integrated with EMRs, and may not be supported by local institutions.

Collaborative Real-Time Mobile TaskList Application
The FastTrack Innovation in Technology team at Boston Children’s Hospital (BCH) developed the first custom build BYOD iOS application, TaskList. It lists shared tasks for a patient and allowed the physician team to use their own devices to check and update tasks through the course of the day and automatically provide live updates to the team members (Figure 1). TaskList is integrated with Epic and Cerner EMRs and complies with Health Insurance Portability and Accountability Act (HIPAA) security and privacy standards. The application was piloted with an inpatient pediatric ward resident team. Using usability survey, we found the application is easy and convenient to use and the team found it is helpful in structuring discussion and task management for patients.

Six Important Characteristics for Patient Hand-Off Application in Inpatient Hospital Settings
Before and during the development of TaskList, we identified a list of fundamental and important requirements for patient hand-off application through a review of literature studies, discussion with external security experts, and reviews with Information Systems Department of BCH. Description and implementation on TaskList for each requirement are listed in Table 1. We used these requirements to build the TaskList application.

Conclusion
We developed and piloted an application designed to facilitate collaborative work and learning and hand-off among physicians using their own devices. We identified that Secure, Integrated, Real Time, Mobile, Collaborative, and Usable are six important and fundamental requirements needed to implement such an application.

References
Figure 1. TaskList user interfaces for patient list, patient detail and task creation, and managing task (update and close).

Table 1. Six Important Characteristics with Description and Implementation on the TaskList application.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Description</th>
<th>Implementation on TaskList</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secure</td>
<td>The system should protect the confidentiality and privacy of patient data. The protection should be applied at the start of the data collection, storing processes, and retrieving back processes.</td>
<td>TaskList uses three layers of security on data communication, works on internal network and VPN only, doesn’t work on jail-broken devices, works with minimal cache, has automatic log off for pre-defined length of inactivity, and limits copy data and print screen.</td>
</tr>
<tr>
<td>Integrated</td>
<td>The system should combine data/features residing in different sources and provide users with a unified view of these data/features.</td>
<td>TaskList is integrated with commercial Cerner and Epic electronic medical records (EMR), LDAP hospital user authentication, and the hospital email system.</td>
</tr>
<tr>
<td>Real time</td>
<td>The system should provide the necessary information that users need within milliseconds so that it is virtually available at the time it is needed.</td>
<td>TaskList is able to populate rarely accessed big data such as patient lists in less than 2 seconds and populate other data in less than 500 milliseconds. To reduce access time, we implemented a cache mechanism on the server side.</td>
</tr>
<tr>
<td>Mobile</td>
<td>The system should be able to move easily and freely in tandem with the users.</td>
<td>TaskList is an iOS application that works on mobile devices (iPad, iPhone, and iPod).</td>
</tr>
<tr>
<td>Collaborative</td>
<td>The system should facilitate recursive process where two or more users work together on a shared task.</td>
<td>TaskList facilitates users to update and download patient care task information (status and note) created by other users as well as providing real time notification, if there are updates on the opt-in task.</td>
</tr>
<tr>
<td>Usable</td>
<td>The system should be easy to learn and use.</td>
<td>TaskList uses simple and usable design for its user interface, and uses BCH enterprise authentication with a Personal Identification Number (PIN).</td>
</tr>
</tbody>
</table>
Capturing Preventive Care Services: Comparing Data Obtained from Manual Chart Review, Automated EHR Extraction, and Insurance Claims

Steffani R. Bailey, PhD¹, Megan Hoopes, MPH², Heather Angier, MPH¹, Rachel Gold, PhD, MPH², John Heintzman, MD, MPH¹, Miguel Marino, PhD¹, Jean P. O’Malley, MPH¹, Jennifer E. DeVoe, MD, DPhil¹²

¹Oregon Health & Science University, Portland, OR; ²OCHIN, Inc., Portland, OR; ³Kaiser Permanente Northwest Center for Health Research, Portland, OR

Introduction

Historically, healthcare quality measures were often derived from: 1) administrative health insurance claims, which fail to capture unbilled services; or 2) manual review of patients’ charts, which is time-consuming, expensive, and often completed on only small subsets of patients. Automated extraction of EHR data presents an efficient, scalable method to assess clinical care quality in large groups of patients. Published comparisons between automated extraction of EHR data and manual chart review have shown mixed results. To further this previous work, we compared automated EHR data extraction to manual chart review on six adult preventive care services that are often performed in the primary care setting (i.e., no referral needed). We also compared Medicaid claims data to manual chart review on the same subset of patients.

Methods

We obtained patient-level data from a linked, hosted EHR serving the OCHIN network of >300 primary care clinics in 16 states. The population was 13,101 adult (ages 19-64), continuously insured Medicaid recipients in Oregon with ≥1 primary care office visit at a study site in 2011. Using automated queries from the EHR, we extracted data on six preventive service metrics ordered or received in 2011: body mass index (BMI), blood pressure, and smoking status assessment; cervical cancer screening; chlamydia screening; and cholesterol testing. EHR data were linked to administrative Medicaid claims data on a unique identification number. From this population, we randomly sampled 150 patients that met age, gender, and clinical eligibility for each screening measure (e.g., cervical cancer screening among women ages 19-64 without total hysterectomy). Manual chart review was conducted by a trained staff member. Prior to completing the chart review, the primary reviewer and a physician-researcher independently reviewed a subset of charts and inter-rater reliabilities were computed. Using standard measures of correspondence (e.g., kappa statistic, sensitivity, specificity), we compared agreement between manual chart review (considered the ‘gold standard’) and the other two data sources on patient-level receipt (yes/no) of each measure.

Results

The patient population from which our subsamples were drawn was primarily female (66%), non-Hispanic white (68%), and low income (92% had family incomes ≤138% of federal poverty level). The mean age was 41 years. Inter-rater reliability was high, with kappa statistics ranging from 0.7-1.0. Automatic data extraction agreed perfectly with manual chart review for BMI and blood pressure assessment and demonstrated very high (>90%) agreement on other measures. Chlamydia screening was captured nearly equally well using the automated method (kappa=.97, sensitivity=.96, specificity=.98), while the assessment of smoking status showed lower agreement (kappa=.66, sensitivity=.97, specificity=.67). Agreement between claims data and manual chart review were low for blood pressure, BMI, and smoking, services that typically do not generate a claim. There was moderate agreement on chlamydia (kappa=.59), cervical cancer (kappa=.67), and cholesterol (kappa=.76) screenings. Agreement between claims data and manual chart review data was lower for all measures compared to agreement between the manual chart review and automated EHR data extraction.

Discussion

Healthcare organizations need accurate methods for evaluating and reporting the quality of care they provide to patients. EHRs are increasingly being used to evaluate quality of care; thus, it is necessary to assess the validity of these data for this purpose. In this study, we found very high to perfect agreement between automated EHR data extraction and manual review of the patient chart, while claims data largely failed to capture unbilled preventive care services typically received in the clinic. These findings suggest that automated EHR extraction is a valid method for capturing preventive service care quality, with advantages over claims data.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Agreement</th>
<th>Kappa</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blood pressure</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual EHR vs. auto EHR</td>
<td>100%</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Manual EHR vs. claims</td>
<td>0.1%</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>--</td>
<td>.007</td>
</tr>
<tr>
<td><strong>BMI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual EHR vs. auto EHR</td>
<td>100%</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Manual EHR vs. claims</td>
<td>12.7%</td>
<td>.002</td>
<td>.008</td>
<td>1.00</td>
<td>1.00</td>
<td>.12</td>
</tr>
<tr>
<td><strong>Smoking status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual EHR vs. auto EHR</td>
<td>94.0%</td>
<td>.66</td>
<td>.97</td>
<td>.67</td>
<td>.96</td>
<td>.71</td>
</tr>
<tr>
<td>Manual EHR vs. claims</td>
<td>10.0%</td>
<td>--</td>
<td>--</td>
<td>1.00</td>
<td>--</td>
<td>.10</td>
</tr>
<tr>
<td><strong>Cholesterol</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual EHR vs. auto EHR</td>
<td>93.7%</td>
<td>.87</td>
<td>.88</td>
<td>.98</td>
<td>.97</td>
<td>.92</td>
</tr>
<tr>
<td>Manual EHR vs. claims</td>
<td>88.6%</td>
<td>.76</td>
<td>.88</td>
<td>.89</td>
<td>.85</td>
<td>.91</td>
</tr>
<tr>
<td><strong>Chlamydia screening</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual EHR vs. auto EHR</td>
<td>98.7%</td>
<td>.97</td>
<td>1.00</td>
<td>.98</td>
<td>.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Manual EHR vs. claims</td>
<td>79.2%</td>
<td>.59</td>
<td>.92</td>
<td>.71</td>
<td>.68</td>
<td>.93</td>
</tr>
<tr>
<td><strong>Cervical cancer screening</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual EHR vs. auto EHR</td>
<td>93.2%</td>
<td>.83</td>
<td>.81</td>
<td>.98</td>
<td>.95</td>
<td>.93</td>
</tr>
<tr>
<td>Manual EHR vs. claims</td>
<td>86.4%</td>
<td>.67</td>
<td>.77</td>
<td>.90</td>
<td>.77</td>
<td>.90</td>
</tr>
</tbody>
</table>

EHR = electronic health record; PPV = positive predictive value; NPV = negative predictive value; BMI = body mass index
-- statistic can’t be computed due to zero cell count in claims data
1 Males and females ages 20-64; cholesterol screening includes LDL, HDL, total cholesterol, and triglycerides
2 Sexually active females ages 19-24
3 Females ages 19-64 with no history of hysterectomy
Medication Compliance in Pediatric Inpatients – What are we missing?

Haresh L. Bhatia a, MS, Neal R. Patel a,b,c, MD, MPH, Catherine H. Ivory d, PhD, RN, BC,
Phillip W. Stewart e, DPh, Kim M. Unertl a, PhD, Christoph U. Lehmanna,b, MD, FAAP, FACMI

a Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN
b Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, TN
c Informatics Center, Vanderbilt University School of Medicine, Nashville, TN
d Vanderbilt School of Nursing, Nashville, TN
e Informatics Clinical Systems, Vanderbilt University Medical Center, Nashville, TN

Introduction

Healthcare providers prescribe medications with the intention to treat or prevent disease. However, according to the World Health Organization, only half of the patients in developed countries adhere to their prescribed medication regimen.

Medication management for pediatric patients poses higher safety risks because children are less able to tolerate errors due to limited internal reserves. It is important that medications for a pediatric patient are administered as intended by the provider. A PubMed search for (“Inpatient” [Mesh]) and (“Medication Compliance” [Major]) did not return any studies conducted on medication compliance for a pediatric inpatient cohort. We studied the compliance rate for pediatric inpatients concentrating on medications not administered to the pediatric inpatient population of the Vanderbilt Children’s Hospital (VCH), which is part of the Vanderbilt University Medical Center (VUMC).

METHODS:

With the IRB approval, we collected order and administration data for pediatric inpatients from the Enterprise Data Warehouse (EDW) at VUMC. We focused only on the medication orders and excluded the orders for which the administration was not mandatory (e.g., PRN orders), was not routinely recorded as medication (e.g., intravenous fluids), or those for discontinuation of medications. We identified the orders without a corresponding administration record as ‘missed-admin’. The set of orders for which the corresponding administration records suggested non-administration was labeled “non-admin” set. Nurses at VUMC must indicate a REASON if a medication was not administered or delayed. We also analyzed the ‘missed-admin’ & ‘non-admin’ sets with respect to the medication class.

RESULTS:

We isolated 1,570,994 medication orders, spanning 42 months (July 1, 2010 through December 31, 2013), corresponding to 117,188 distinct pediatric inpatients. There were 3,742,013 matching administration records for these orders. Only 596 distinct medication order records did not have a valid matching administration record (missed-admin) (Table 1).

For the non-admin, we grouped the REASON-CODEs into three categories: “Administered per protocol” (medication was given but in a modified manner – e.g., when the patient was off the unit), “D/C per protocol” (circumstances like being in surgery prevented administration), and “Not administered”. There were 101,688 administration records in the “Non Administered” category, corresponding to 41,099 medication orders. The top two medication classes comprising the missed and non-administered doses were “Alimentary tract and metabolism drugs” (e.g., Docusate Sodium, Creon) and “Nervous system drugs” (e.g., Morphine, Fentanyl)

DISCUSSION and CONCLUSION:

The medication process (ordering, dispensing, and administration) in pediatric care is complex. Health information technology can help to measure the quality status of the medication process at various levels. To our knowledge no studies have ever evaluated medication administration compliance in a children’s hospital.

The small number of missing administration records (596 of the total 1,570,994 orders – < 0.05%) demonstrates the robustness of the electronic Medication Administration Records (eMAR), which prompts nurses to document every scheduled dose. This small gap could reflect truly not administered and undocumented doses. The non-admin records corresponded to < 3% (41,099) of the distinct orders. The absolute number may seem concerning, however breakdown of reasons demonstrates that the majority were compliant.

Our findings suggest that the medication process in pediatric practice is rational and reliable.
Table 1: Overview of Pediatric Medication Administration Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinct Pediatric Inpatients</td>
<td>117,188</td>
</tr>
<tr>
<td>Distinct number of Medication Orders</td>
<td>1,570,994</td>
</tr>
<tr>
<td>Matching Administration Records</td>
<td>3,742,013</td>
</tr>
<tr>
<td>Missing Administration Records</td>
<td>596</td>
</tr>
<tr>
<td>Administration records with REASON-CODE indicating non-administration</td>
<td>101,688</td>
</tr>
<tr>
<td>Distinct medication order records corresponding to the administration with a REASON-CODE indicating non-administration</td>
<td>41,099</td>
</tr>
</tbody>
</table>

Table 2: The Medication Administration Records with a Non-Zero REASON-CODE (indicating Non-Administration or Deviation from the Stipulations of the Order)

<table>
<thead>
<tr>
<th>Description</th>
<th>Category</th>
<th>Administration Count</th>
<th>Proportion (%)</th>
<th>Category Prop. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Med Discontinued</td>
<td>Permitted Deviations</td>
<td>18,429</td>
<td>7.07</td>
<td></td>
</tr>
<tr>
<td>Per MD order</td>
<td></td>
<td>10,919</td>
<td>4.19</td>
<td></td>
</tr>
<tr>
<td>NPO</td>
<td></td>
<td>10,782</td>
<td>4.14</td>
<td></td>
</tr>
<tr>
<td>Per Parameter</td>
<td></td>
<td>3,631</td>
<td>1.39</td>
<td></td>
</tr>
<tr>
<td>Sleeping</td>
<td></td>
<td>3,398</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td>No insulin required</td>
<td></td>
<td>2,288</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>No IV access</td>
<td></td>
<td>2,176</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>Nauseated/Vomiting</td>
<td></td>
<td>1,973</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>Sedated</td>
<td></td>
<td>272</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>Unable to swallow</td>
<td></td>
<td>216</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Not on Unit</td>
<td></td>
<td>121</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>VPH Partial Patient</td>
<td></td>
<td>66</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Other- annotate</td>
<td>Not Administered</td>
<td>39,111</td>
<td>15.00</td>
<td></td>
</tr>
<tr>
<td>Patient / Family refused</td>
<td></td>
<td>29,587</td>
<td>11.35</td>
<td></td>
</tr>
<tr>
<td>Acknowledged</td>
<td></td>
<td>573</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>REASON NOT GIVEN</td>
<td></td>
<td>290</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Med not available</td>
<td></td>
<td>53</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>260,665</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

REFERENCES:

The Hidden Life of Nurses’ Cognitive Artifacts

Jacquelyn W. Blaz, MS1, Alexa K. Doig, PhD, RN1,
Kristin G. Cloyes, PhD, MN, RN1, Nancy Staggers, PhD, RN, FAAN1
1University of Utah, Salt Lake City, UT

Introduction

Standardizing handoffs is recommended to improve communication, with electronic tools as the primary approach. However, nurses continue to rely on paper-based tools to support handoffs, indicating a deficiency in available electronic versions. Nurses’ paper tools have been described as cognitive artifacts1,2—tools that provide cognitive support by offloading part of the cognitive work required for a task from the human mind to an external object. Characterization of nurses’ cognitive artifacts as merely handoff tools may be a limited view of the purposes these objects serve and the processes leading to their construction. Previous research hints at uses beyond handoff, but little work has been done to examine uses outside the context of handoff1,2. Electronic tools developed to support nursing handoff that disregard nurses’ cognitive needs beyond handoff may be doomed to be underutilized. The purpose of this research was to develop a deep understanding of the use and production of nurses’ paper cognitive artifacts in a medical oncology unit.

Methods

A grounded theory approach3 was used to explore nurses’ handoff artifacts. Seventy-three hours of field observations in a medical oncology unit in a cancer specialty hospital led to 13 purposively sampled nurses. These nurses were shadowed for an entire shift and interviewed after giving handoff to the next shift. The data corpus included images of paper cognitive artifacts, transcribed interviews, field notes, and analytic memos. Consistent with grounded theory techniques, the data were coded and collected into categories of similar ideas. Concepts emerged from further analysis and interpretation of codes and categories.

Results

Nurses' paper-based artifacts are deemed by nurses to be personal, dynamic, living objects that traverse a life cycle during each shift and evolve during the course of a nurse's career. The life cycle has four phases. Origin occurs with the creation of a new artifact at the beginning of the shift; the second phase is the use of the artifact throughout an entire shift; reproduction is the transfer of information by a nurse from an old artifact to a new one; and the final phase occurs with the destruction of an artifact when it is considered no longer useful.

Nurses' paper artifacts experience an evolutionary process over nurses’ careers. Evolution in a nurse's individually styled, paper artifact is triggered by a change in the nurse's environment that reshapes cognitive needs. If an artifact is not able to provide cognitive support in a new environment, it is modified into (adapted), or abandoned for (made extinct), a different format that will provide the necessary support. For example, a newly hired nurse may create an area to record information related to chemotherapy regimen to an existing paper tool s/he created for use on a general medical unit. With each new nursing shift and artifact life cycle, a nurse may either abandon or modify the artifact until a new design solidifies that is “good enough” for the nurse's cognitive needs.

Discussion

The two processes of nurses’ cognitive artifacts--life cycle and evolution--point to the need for extreme flexibility in future designs of standardized handoff tools, especially electronic versions. Individual cognitive needs may not always be in congruence with group needs, thus a ‘one-size-fits-all’ approach to design for an electronic handoff tool is not recommended. Not only must electronic tools have the ability to be updated quickly and easily by the user during a shift, but the overall display design must also be changeable by an individual nurse for the immediate context of work, the patient's condition and to accommodate changes in cognitive work over time. This could be addressed with periodic evaluations of how an electronic form is working for nurses on a unit, especially following policy changes affecting nursing workflow or by utilizing a display design consisting of modules of related information placed according to the preference of individual nurses.
References


Real Time Active Learning Study for Clinical Named Entity Recognition

Yukun Chen, MS1; Sungrim Moon, PhD2; Thomas A. Lasko, MD, PhD1; Qiaozhu Mei, PhD3; Jingqi Wang, MS2; Trevor Cohen, MBChB, PhD2; Qingxia Chen, PhD4,5; Joshua C. Denny, MD, MS1;5; Hua Xu, PhD1,6

Department of 1Biomedical Informatics, 4Biostatistics, and 6Medicine, Vanderbilt University, Nashville, TN; 2School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX; 3School of Information and Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI.

Introduction: Active learning techniques are to select samples for annotation based on an estimation of their utility for predictive modeling, rather than at random. They have been studied in biomedical natural language (NLP) tasks, such as assertion classification for clinical concepts1, word sense disambiguation in biomedical literature2, and phenotyping from electronic health records.3 These studies have demonstrated that active learning can help achieve high-quality supervised classification models with reduced annotation cost. Most of these simulated studies assume that annotation cost for each sample was identical. In reality, however, annotation cost (i.e. the time required by an annotator) can be very different from one sample to another. In this study, we evaluated the real-time performance of active learning versus passive learning for a clinical named entity recognition (NER) task. Active LEARNER, an active learning enabled annotation system for NER, was developed to support the user study. We collected the annotation data from two nurses and generated learning curves, analyzing F-measure of NER models as a function of real annotation time, for both active and passive learning. We find that active learning is not guaranteed to reduce actual annotation time to achieve a model with desired performance, when compared to passive learning.

Methods: The clinical NER task in this study was to extract problem, treatment, and lab test concepts from clinical notes. We used the 2010 i2b2 NLP challenge dataset to generate the querying pool and test set for the NER model assessment, and to evaluate the annotation quality. Our study consists of three parts: (1) development of Active LEARNER, (2) user study design, and (3) evaluation:

(1) We propose Active LEARNER, which uses a preprocessed unlabeled corpus as the input, actively queries sentences and presents them to the interface for annotation, and generates NER models in real time. There are three major components of the system: annotation interface, learning engine, and querying engine. For the annotation component, the system embeds “brat”, a rapid annotation tool2, as an interface for entity tagging. The learning engine is based on the CRF algorithm implemented by CRF++.7 The querying engine implements our newly developed active learning algorithm called CAUSE (Clustering and Uncertainty Sampling Engine), which considers both uncertainty and diversity of sentences in the corpus. Figure 1 presents the multi-thread workflow design of Active LEARNER. The learning process, which mainly includes CRF model encoding and unlabeled sentences ranking, is parallelized with the annotation process. We upload a video online6 to demonstrate the annotation process using our interface.

(2) In the user study, CAUSE and RANDOM sampling methods represent active learning and passive learning, respectively. We recruited two nurses to participate in the user study of Active LEARNER. To ensure that the results from the two methods are comparable, users need to be well trained and perform consistently in both experiments. Therefore, the user study includes two parts: (a) annotation training, and (b) the main study. Annotation training includes review of annotation guidelines, a review of some sentence-by-sentence annotations, and annotation practice. Once users had completed the training, they started the main study for each method with section management.

(3) We analyzed users’ responses over time on both methods in the main study. The evaluation metrics included words/sentences per minute (average sentence length), entity annotations per minute (annotation speed), words per minute (reading speed), and F-measure (annotation quality, based on i2b2 gold standard). We plotted the annotation curves for some of these responses (e.g. entities versus annotation time and words versus annotation time) for the two users across the two methods. We also show the learning curves of NER modeling (e.g. F-measure versus annotation time) across the two methods and the two users. A survey of the two users was done to gain their general impression on the difference between the two methods from the perspectives of annotation difficulty, clinical relevance of sentences, diversity of sentences, etc.

Results: Figure 2 shows the annotation curves of reading speed and annotation speed by RANDOM and CAUSE for user 1. Figure 3 shows the same types of curves for user 2. In terms of the overall annotation speed, user 1 annotated 7.83 and 7.71 entities per minute by RANDOM and CAUSE, respectively, while user 2 annotated 7.28 and 7.94 by RANDOM and CAUSE, respectively. Both users achieved F-measures ranging from 0.79-0.81 for overall annotation quality for each method. Figure 4 presents the learning curves of NER modeling by the two methods from the two users. The learning curves show that RANDOM performed slightly better than CAUSE (by eye) for user 1, while CAUSE slightly outperformed RANDOM for user 2.

Discussion: Active learning selected longer and more clinically relevant sentences compared to random sampling. However, annotating the actively selected sentences was more difficult (by survey) and took more time. While previous studies have found increased NER performance per training sentence using active learning, we found no reliable increase in performance per annotation time, suggesting that the increased information content of actively selected sentences is strongly offset by the increased time required to annotate them. Moreover, the higher annotation speed from users induced better learning curves. Interestingly, the overall reading speed is also higher for randomly sampled sentences (62-68 words per minute) compared to actively selected sentences (53-54 words per minute). Our results demonstrate the need to benchmark active learning algorithms using real-world practical measures (such as user time) in addition to more theoretical measures (such as the number of training sentences). The next phase of our work will include improving our active learning algorithms against the practical measures.
Figure 1. Workflow of Active LEARNER

Figure 2. Annotation curves of average sentence length, reading speed, and annotation speed by RANDOM and CAUSE from user 1

Figure 3. Annotation curves of average sentence length, reading speed, and annotation speed by RANDOM and CAUSE from user 2

Figure 4. Learning curves of F-measure versus minutes by RANDOM and CAUSE from user 1 and 2

Reference
A REAL TIME ELECTRONIC REGISTRY AS A KEY INTERVENTION TO REDUCE TREATMENT DISPARITIES IN EARLY STAGE, NON-SMALL CELL LUNG CANCER: PRELIMINARY RESULTS

Samuel Cykert, Paul Walker, Peggye Dilworth-Anderson, Maria Cirino-Marcano, Lloyd Edwards

INSTITUTIONS:
1. General Medicine and Clinical Epidemiology, University of North Carolina, Chapel Hill, NC, United States.
2. Leo Jenkins Cancer Center, Brody School of Medicine, Greenville, NC, United States.
3. Gillings School of Global Public Health, University of North Carolina-Chapel Hill, Chapel Hill, NC, United States.
4. Pulmonary and Critical Care, University of South Carolina School of Medicine, Columbia, SC, United States.

INTRODUCTION: African-Americans (AA) with lung cancer experience a higher annual death rate compared to Whites (W) with AA men particularly affected (78.5 vs 65.7 deaths per 100,000 population). Despite this risk, Bach and others have shown that over the last 3 decades surgical rates for AA patients with potentially curable stage 1 and non-small cell cancer lag behind surgical rates for W patients. Although stereotactic radiation cures some patients with smaller lesions, lung cancer surgery remains the treatment of choice according to National Comprehensive Cancer Network guidelines. These guidelines were further supported by Gray et al. who demonstrated in a small areas analysis that regions of the U.S. with high surgical rates (>79%) have less lung cancer mortality than areas with lower surgical rates even considering high risk patients. These results make the treatment difference all the more remarkable. Because of the surgical gap and associated mortality differences, we launched a multifaceted interventional trial designed to optimize lung surgery rates for all patients with early stage lung cancer reduce disparities associated with AA race. We report the early results of this effort.

METHODS: We first performed a 3-year retrospective chart review for all patients with biopsy proven lung cancer at 3 academic institutions to establish baseline surgical rates according to demographic features, including race, and severity of illness. Starting September of 2012, we prospectively recruited patients with Stage 1 or 2 non-small cell lung cancer who were identified at initial visit scheduled with a pulmonologist, surgeon, or at a multidisciplinary lung cancer clinic at the 3 institutions where the retrospective chart reviews were performed. Patients were approached if he/she had either biopsy proven disease or qualified through a Bayesian probability algorithm based on the results of a recent chest computerized tomography scan. This report includes patients who reached study endpoints (treatment given or no treatment administered within 4 months of enrollment) by January of 2015. Immediately after enrollment, we then applied the multipronged intervention that used digital systems, quality improvement techniques and an enhanced care team at each study site. The specific intervention components are as follows:

(1) a real time electronic registry that provides an electronic warning within the registry if either a patient misses an appointment or doesn’t reach a designated milestone in care. Data in the registry for appointments, procedures, and diagnostic tests are derived from electronic health records. Warnings were either noted by the enhanced care team or delivered to members of the usual care team.

(2) quarterly feedback of surgical rates by race and co-morbid illness to providers and other cancer care personnel in the participating centers. This data reporting creates an environment of transparency.

(3) a specially trained navigator who develops patient rapport early then reengages patients immediately after any lack of adherence to appointments or recommended care. If a real time registry warning occurs because a care milestone has not been reached, the navigator is also empowered to address this circumstance with the care team.

Given that electronic health records are now ubiquitous, interventions (1) and (2) were used for all patients and compared to retrospective controls. Consented patients were randomly assigned (3) to special navigation or usual care.

RESULTS: Baseline surgical rates from the retrospective analysis of 714 early stage, non-small cell patients were 69% for W and 66% for AA patients. When controlling for comorbidities, COPD, age, and other demographic data, the Odds Ratio for surgery for AA compared to W lung cancer patients was 0.64 (95% CI 0.43, 0.96). To date 157
patients have prospectively reached study endpoints including 54 AA patients (34%). The mean age for this study group is 65 years; 53% are female. 49% were assigned to special navigation. Preliminary findings show an overall surgical rate of 81% (82% W, 78% AA, p = 0.5). In multivariate analysis, using demographic variables, comorbid illness, pulmonary function, and patient perceptions of communication as independent variables, age > 70 (OR 0.18, 95% CI 0.38, 0.83) and a forced expiratory volume < 40% of predicted (OR 0.04, 95% CI .001, 0.19 ) were associated with lower surgical rates. Notably, AA race and medical comorbidity scores did not affect surgery.

CONCLUSION: Early results from a multifaceted intervention designed to optimize lung cancer surgery and narrow the surgical gap between AA and W patients appear promising.

References:


Improving Care Team Communication: Early Experience at Implementing a Patient-centered Microblog

Anuj K Dalal, MD,1,2 Jeffrey L Schnipper, MD, MPH,1,2 Anthony Massaro, MD,1,2 Kelly McNally,1 Patricia Dykes, RN, DNSc,1,2 and David W Bates, MD, MSc,1,2

1Brigham and Women’s Hospital, Boston, MA, 2Harvard Medical School, Boston, MA

Introduction: Care team communication in acute care settings is fragmented, inefficient, and difficult.1,2 Inpatient and ambulatory providers frequently communicate asynchronously via alpha-numeric pagers, network email, and increasingly, unsecure texting to discuss and coordinate the patient’s plan of care. We engaged institutional stakeholders in the design and development of a patient-centered microblog. The patient-centered microblog aims to facilitate seamless care team communication by accurately and reliably identifying care team members and directing providers to a single conversation thread on which to view and contribute to the patient’s plan of care discussion. The platform supports access from web-enabled desktop computers and mobile devices and is a component of two institutional initiatives–PROSPECT and PCORI Transitions–that serve to promote patient-centered care team communication during hospitalization. The purpose of this study is to describe our experience at implementing the patient-centered microblog in the acute care setting.

Methods: This study took place at a large academic medical center in Boston, Massachusetts. The system was implemented on 2 MICU, 2 Oncology, and 2 General Medicine units. We presented a structured overview to attendings, housestaff, physician assistants, fellows, nurses, social workers, care coordinators, primary care physicians, and case managers throughout the study. During each session we reviewed secure log-on, core functionality (real-time care team member identification, read-receipts, persistence of messages on threads, admission and message notifications, etc.), workflow integration, and mobile access. We measured messaging activity over a 2-month period after core functionality was fully implemented. We recorded all feedback. We conducted three, 1-hour focus groups with 4-6 providers (housestaff, physician assistants, and nurses who had worked on intervention units). We conducted ½-hour interviews with intervention unit leadership (3 physicians, 3 nurse managers). We addressed the following categories: core functionality, actual clinical use, notifications and alert fatigue, barriers to use, and desired enhancements. We used content analysis methods to interpret descriptive data from focus groups and interviews. We used a 2-person consensus approach to identify major themes.

Results: In general (Figure 1), the volume of weekly messages increased over time (Figure 1a), but varied by type of clinical unit (Figure 1b). Of 11,277 message notifications sent, 8308 (74%) were viewed by the notified recipient. See Table 1 for descriptive feedback and derived themes. Overall, core functionality was favorably received and actual clinical use matched intended use-cases. Concerns regarding alert fatigue were present but reflected variable understanding of functionality that was incorporated to minimize excessive notification. Barriers to using the system–namely, competing messaging systems (e.g., email) and maintaining accuracy of care team member role assignments from scheduling systems–did not seem insurmountable. Strategies to overcome these barriers included active participation by intervention unit leadership and alignment with institutional culture change initiatives. Requests for enhancements were a promising indicator of the system’s potential. Finally, providers were willing to adopt and enhance the tool, even among those who were initially skeptical (Box 1).

Conclusion: Overall, our experience at implementing a patient-centered microblog to improve care team communication has been positive. Although intervention unit leadership supports its use, overall adoption has been variable so far. In instances when the system has been used, the perception of its value to coordinating the plan of care among care team members has been favorable. Our next steps are to continue to promote adoption and consider how to sustain and expand scope of use as our institution transitions to a vendor-based electronic health record.

Acknowledgements: The Brigham and Women’s Hospital PROSPECT initiative is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation.

References:
Figure 1: Weekly Care Team Messaging Activity via Patient-centered Microblog

Table 1: Major Themes from Focus Groups and Interviews after Implementation

<table>
<thead>
<tr>
<th>Topic</th>
<th>Feedback from Focus Groups &amp; Interviews</th>
<th>Derived Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functionality</td>
<td>Care team is overall accurate and up-to-date</td>
<td>System functionality has been favorably perceived, serving to keep the care team on the same-page effectively and efficiently as intended.</td>
</tr>
<tr>
<td></td>
<td>Read receipts helps sender know when message was viewed by recipient</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A microblog format is ideal because messages remain in one virtual place ( persistence ); no need to forward messages to new care team members as they come on shift, when new members are added to the care team, or when patient transitions to a new team</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Solves the problem of keeping shift-based providers (nurses, responding clinicians) up-to-date → message notifications are automatically routed to correct providers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Message counts on clinical applications helps to identify new message activity</td>
<td></td>
</tr>
<tr>
<td>Actual Clinical Uses</td>
<td>Mostly for non-urgent clinical messaging and updates</td>
<td>Perceptions of actual use is consistent with intended use-cases, however original scope may be too restrictive.</td>
</tr>
<tr>
<td></td>
<td>Ambulatory provider input into plan of care (medical problems, psychosocial issues)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multidisciplinary care coordination and goals of care planning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Limited to single, broad topic; no ability to conduct specific topic-based conversations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Limited to specific inpatient units; limited ability to use post-discharge</td>
<td></td>
</tr>
<tr>
<td>Notifications and alert fatigue</td>
<td>Volume of email notifications is generally acceptable; some think it can be excessive</td>
<td>Concerns regarding alert fatigue exist but reflect variable understanding of functionality that we have incorporated to minimize excessive notifications.</td>
</tr>
<tr>
<td></td>
<td>Notifications are non-descriptive and do not include identifying patient information, rendering them less useful</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lack of understanding of message persistence: care team members can view messages “on-demand” regardless of whether they were notified (e.g., nurse can view entire conversation thread, including prior messages sent to others when coming on shift)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lack of understanding of how to effectively configure mobile app notifications to manage alert fatigue → push notifications are often disabled on mobile devices</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unaware that email notifications are suppressed when recipient receives a push notification on mobile device prompting them to view the message on the mobile app</td>
<td></td>
</tr>
<tr>
<td>Barriers to use</td>
<td>Expectations for use by unit based providers is variable, and largely responsible for the observed activity on specific clinical units (MICU) → active participation and engagement of clinical unit leadership is instrumental to drive adoption</td>
<td>Barriers to using system are not insurmountable, and mostly related to efforts at engaging leadership, ensuring accuracy of care team member role assignments from paging directories and scheduling databases, and sustaining institutional culture change initiative.</td>
</tr>
<tr>
<td></td>
<td>Use of competing messaging systems (email), especially by non-unit based providers (ambulatory providers, consultants), increases perception of messaging silos</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Care team role assignments are sometimes incorrect because scheduling databases are not always up-to-date and not use of role assignment functionality is not enforced administratively by the institution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial log-on to mobile application is time-consuming → must enter username and password and answer security questions prior to setting a 4 digit pin</td>
<td></td>
</tr>
<tr>
<td>Desired Enhancements</td>
<td>Ability to initiate multiple topic threads should increase adoption and use</td>
<td>Ongoing requests for enhancements are a promising indicator of the system’s potential.</td>
</tr>
<tr>
<td></td>
<td>Incorporate functionality to remove oneself from notification recipient list to reduce alert fatigue when a provider no longer needs to actively participate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Allow protected health information and message previews in email notifications</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Include non-network providers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Incorporate ability to search conversation thread and filter messages</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ability to upload photos and attachments securely</td>
<td></td>
</tr>
</tbody>
</table>

Box 1: Comments and Feedback from Providers

“I was skeptical about using this tool at first, but after using it, I am a convert. It keeps all the messages about the patient in one place, not interspersed with email in my inbox. I changed the care that my patient received.” – A hospitalist attending

“I think it has a lot of promise. Can see it being very nice for keeping PCPs up-to-date with the smaller details… It is easy enough to read a discharge summary, but some of the details get lost.” – A primary care physician

“I was a consultant but rotated off after a week….it would be nice to have a button on the bottom of the message saying ‘Remove Me’” – A subspecialty consultant

23
Using a Software Program to Support Shared Decision Making about Participation in Breast Cancer Clinical Trials

Prudence Dalrymple, PhD, MS, Lisl Zach, PhD, Michelle Rogers, PhD
Amy Leader, DrPH, MPH, Tiffany Avery, MD, MPH, Anna Quinn, MPH, Anett Petrich, RN, Massimo Cristofanilli, MD, Russell Schilder, MD, Ronald Myers, PhD
1College of Computing & Informatics, Drexel University, Philadelphia, PA;
2Department of Medical Oncology, Thomas Jefferson University, Philadelphia, PA

Introduction

In the United States, over 230,000 women will be diagnosed this year with invasive breast cancer, resulting in about 40,000 deaths. Although treatment advances depend on clinical trials, numerous barriers to participation exist and interventions to promote informed, shared decision making may increase trial participation. We report on the use of an online decision support tool based on Analytic Hierarchical Processing to help breast cancer patients and their physicians make more informed decisions about trial participation.

Methods

Prior to an appointment with their oncologist, eligible patients meet a health educator and complete a baseline knowledge survey about clinical trials. Together, they review a one-page grid comparing risks and benefits of trial participation; they then use the online Decision Counseling Program (DCP) to elicit the patient’s preferences about trial participation. The patient expresses her top 3 reasons “pro” and “con” participation and rates their importance. The factors and their rating are entered into the DCP; the algorithm calculates a preference score from 0.000 to 1.000 and then produces a one-page summary report (SR) (see Figure 1). The patient takes the SR to her doctor’s appointment where trial participation is discussed. At 30 days, a telephone survey assesses the patient’s enrollment status, satisfaction with decision-making, and perceptions of the decision counseling session. The clinician documents trial enrollment.

Results

To date, 19 patients have enrolled in the study. “Pro” factors (36) included learning of new treatment options and believing that participation could benefit others. “Con” factors (19) were fear of unknown side effects and concern about managing trial logistics. The mean decision preference score was 0.6742 (SD=0.1465). Categorical assessment showed that one patient did not favor participation, 3 were equivocal and 15 favored participation. Eleven of the 19 patients have completed an endpoint survey at 30 days (5 were lost to follow-up and 3 are not yet due). Patient knowledge increased an average of 1 point, from 7 correct answers at baseline to 8 correct at endpoint. Ninety day chart audits indicate that trial enrollment has increased from an historical rate of 10% to a rate of 30%.

Discussion

The decision counseling process--preference clarification plus education--prepares patients and clinicians for shared decision-making by presenting the patient’s current attitudes and beliefs in a consistent format in the clinical workflow. These features address issues that have been problematic in other studies. Although this study focuses on cancer, the DCP can be adapted to other clinical situations. Furthermore, the DCP aggregates patient responses in an online database that allows researchers to identify the most frequent and important factors; these data can be extracted and analyzed to determine patterns in subpopulations. Additional research is needed to examine the characteristics of patient-provider communication to examine its effect on shared decision-making.

Conclusion

By supporting the patient through education and clarification of preferences and values, the DCP provides a supportive tool that appears to reduce decisional conflict and increase satisfaction while advancing the twin goals of shared decision-making and patient empowerment.
Division of Population Science
Decision Counseling Report: Likelihood to join a breast cancer clinical trial

Decision to be Made - Option 1: Likely to join a clinical trial for breast cancer at this time or Option 2: Not likely to join a clinical trial for breast cancer at this time.

Session results indicate that your position is neutral.

<table>
<thead>
<tr>
<th>Option 1</th>
<th>0.341</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Option 2</td>
<td>0.459</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Top Decision Factors and Direction of Influence:

<table>
<thead>
<tr>
<th>Factor</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>I want to help find a treatment that will help cure me and others</td>
<td>Favors Option 1</td>
</tr>
<tr>
<td>I worry about dealing with additional side effects</td>
<td>Favors Option 2</td>
</tr>
</tbody>
</table>

Comments:

1. I would like more information about the specific trials that are open to me.

I understand and agree with the Decision Counseling Report results shown above.

Participant ID: BCCT0003094 example
Participant First Name: Example
Participant Last Name: Example
Participant Signature: ____________________________

Date: 06/28/2014

References

An EHR-Integrated Shared Decision Making Mobile App for Prostate Cancer Screening

Frank C. Day, MD, MPH, Majid Sarrafzadeh, PhD, Stephanie Smith, MPH, CPHQ, Mohammad Pourhomayoun, PhD, Costas Sideris MSc, Amogh Param MS, Jonathan Ben-Hamou BA, Deidre Keeves, PT, Michael A. Pfeffer, MD, Douglas S. Bell, MD, PhD  University of California, Los Angeles, CA

Introduction

Engaging patients in shared decision making (SDM) is important to optimize patient-centered outcomes and help control health care costs and patients who use SDM decision aids have had superior outcomes. Electronic health record (EHR) patient portals do not enable patients and families to engage in SDM, a difficult, time-consuming process. To do SDM, a patient needs a clear comparison of the likelihoods of benefit and risks of harm associated with each option, and guidance to the choice most congruent with the patient's values and preferences. Quantitative risk comparison is difficult because the concept of probability is hard to understand, particularly for outcomes that are binary in any individual. A large RCT to evaluate a SDM web app (with visual risk comparison diagrams to convey the risk tradeoff inherent to the PSA screening decision for prostate cancer) found patients' ratings of SDM did not differ between groups, but the app did positively affect the quantity and content of SDM discussions that physicians subsequently had with their patients. A mobile app will read EHR data (to generate risk diagrams based on the documented race and family history of prostate cancer), and write data in real-time to the EHR based on a patient's inputs (activation level, updated risk factors, values selections, and decision preference).

Methods

A multidisciplinary team (including patient advisor, health IT, computer science, and vendor technical services) is producing a mobile app/EHR integration with machine learning (ML) analytics. Our timeline (Mar 2014 – Nov 2015) includes the adaptation of a web-based SDM app to a mobile platform, establishing the architecture for secure data exchange that complies with privacy regulations and maintains data integrity, and limited usability testing with live patients. The target population is 55-69 year old men, who will use the app in primary care clinic and at home. Outcome measures will include usability and ML metrics. Multiple ML methods (e.g. support vector machine, decision trees, naive Bayes) will be analyzed for best fit of an expected ~300 dimension space (measured and derived from EHR, location and social media data) hyperplane to predict outcome variables (which patients specify a decision preference, which preference is specified, and whether preferences are congruent with subsequent medical management).

Results

Having committed support from the health system and a dedicated project management expert is key to overcome the numerous challenges with a project of this scope and timeline. Continual emphasis on both improving patient-centered care and maintaining data integrity, security and privacy is critical to maintain the alignment of goals for a diverse set of teams. Constraining the EHR interface to 9 data elements, and adapting a preexistent web app with established efficacy (that requires minimal new content development for a mobile platform) enables the team to focus on the data security, regulatory and policy/governance issues that arise with an early mHealth initiative. IRB approval provides the expectation of outcomes data and can be an effective driver to prioritize the allocation of necessary human resources.

Discussion

Secure bidirectional data exchange with an EHR is feasible and may increase patient engagement and improve patient-centered outcomes (AI–II.B: Clinical decision support and guideline systems). A ML algorithm may find unexpected predictors of optimal SDM outcomes within EHR, location and social media sources (FI–I.B: Social Media and mobile technology (mHealth)).
Figure 1. System architecture

Figure 2. Risk diagrams

References

Polychromatic X-Ray Absorptiometry to Quantify Breast Density Volume, Ratio and their Associated Breast Cancer Risk in Full-Digital Mammography

Luis de Sisternes, PhD, \(^1\) Joseph H. Rothstein, MS, \(^2\) Abra M. Jeffers, \(^3\) Weiva Sieh, MD, PhD, \(^2\) and Daniel L. Rubin, MD, MS, \(^1,4\)

\(^1\) Department of Radiology, \(^2\) Health Research and Policy, \(^3\) Management Science and Engineering, and \(^4\) Medicine, Stanford University, Stanford, CA, USA

1. Introduction

Mammographic density, which refers to breast tissues other than adipose tissue (like epithelial or stromal tissue), is one of the strongest cancer risk factors\(^1\). Several area-based\(^2\) and volume-based\(^3,4\) approaches to quantitate breast density have been widely used. As screen-film mammography is being replaced by full-digital mammography (FFDM), existing methods originally described for digitized film mammograms were adapted to FFDM and many other are emerging given the easier implementation of advanced applications in a digital setting.

The semi-automated, area-based Cumulus approach has been reported to have the strongest cancer risk predictive ability\(^5\) in FFDM images, although it is limited by intra- and inter-reader variability in establishing a threshold for differentiating dense from surrounding fatty tissue. Our aim is to develop a novel, fully automated method to provide direct dense tissue measurements presenting higher association with cancer risk than Cumulus, while also eliminating any subjective judgement in such measurements. Our method is based on polychromatic X-ray absorptiometry (PXA), nonlinear programming, and image processing techniques to estimate dense volume and ratio of dense to adipose tissue on a pixel-by-pixel basis throughout FFDM images. We evaluated its ability to predict breast cancer risk, and compared it to Cumulus dense area and percent density (PD) measurements.

2. Methods

The idea behind our proposed method is based on single energy X-ray absorptiometry\(^6\), but taking into account the fact that the mammogram is produced by a polychromatic X-ray source (we do not assume a monochromatic source). The method does not require a phantom in the acquired images. Following the well-known Beer-Lambert law and assuming that breast is mainly composed of adipose and dense tissue, the raw intensity values recorded at the detector can be expressed in the terms shown in Figure 1. This expression can be solved for the ratio of adipose and dense tissue at each pixel location using a constrained non-linear optimization technique\(^7\), where the energy-dependent attenuation coefficients of each tissue type are known\(^8\), the source spectrum is simulated\(^9\), and breast thickness can be estimated in a similar manner as in de Sisternes et al\(^10\). Two overall measurements were computed from the pixel-by-pixel estimation: (1) absolute PXA dense volume, computed by adding pixel by pixel the result from the multiplication of dense ratio and breast thickness; (2) PXA volumetric percent density (VPD), computed by averaging the pixel-based dense ratio values throughout the mammogram, weighted by breast thickness.

We evaluated our methods in 131 mammograms from unaffected breasts prior to a cancer diagnosis in the contralateral breast (cases) and 239 mammograms from healthy women without breast cancer (controls). Control women were chosen to match the case patients by age and race. Patient demographics are summarized in Table 1. The study protocol was approved by the Stanford University Institutional Review Board. Cumulus dense area and PD values were collected by an expert user of the software for comparison to those produced by our methods.

3. Results

We compared the distribution of patient demographics, Cumulus and PXA measurements in case and control women by computing false discovery rate adjusted q-values (Table 1). Table 2 summarizes the odds ratios (OR) estimated for the two methods as quartiles (defined among controls) and as continuous variables (in SD increments), unadjusted and adjusted for age, BMI, race and menopausal status. We computed the area under the receiver operating characteristic curve (AUC) for each method to compare their ability to discriminate between cases and controls. Overall, slightly higher associations with breast cancer risk were observed for our proposed method.

4. Discussion

Both Cumulus and PXA measurements were positively associated with breast cancer risk, with the strongest continuous association observed for each SD increment of PXA VPD. Women in the top quartile of PXA VPD had 5.30 (1.99 to 14.16) times the risk of women in the bottom quartile. PXA measurements were fully automated and not operator-dependent, eliminating any possible subjective variability. Although a highly-trained reader generated the Cumulus measurements, reader judgement may have contributed to their slightly weaker associations with breast cancer risk. PXA measurements also showed a slightly greater ability to discriminate between cases and controls in terms of AUC. These AUC values were low (0.64 for dense volume and 0.65 VPD), but similar to that previously reported by others, indicating its limited value in individual cancer risk prediction. Therefore, the PXA method presented here appears to be a valid automated alternative to the labor-intensive semi-automated Cumulus approach for quantifying breast density when raw FFDM images are available for analysis, and it also offers the possibility of pixel-by-pixel analysis of volume-based methods. These quantifications, alone or jointly with other risk factors, might be useful to stratify women in the population according to risk for tailored screening or interventions.
Figure 1. Expression for estimating the proportion of adipose tissue and dense tissue in raw FFDM images.

Table 1. Distribution of patient characteristics in case and control women.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Case mean (SD) or % (n=131)</th>
<th>Control mean (SD) or % (n=239)</th>
<th>Total mean (SD) or %</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>54.35 (10.90)</td>
<td>55.03 (11.04)</td>
<td>54.79 (10.98)</td>
<td>0.35</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>26.54 (6.21)</td>
<td>25.64 (5.87)</td>
<td>25.96 (6.00)</td>
<td>0.21</td>
</tr>
<tr>
<td>Race (% Caucasian)</td>
<td>62.83</td>
<td>62.76</td>
<td>62.43</td>
<td>0.44</td>
</tr>
<tr>
<td>Cumulus Dense Area (cm³)</td>
<td>33.68 (19.74)</td>
<td>28.72 (17.41)</td>
<td>30.47 (18.39)</td>
<td>0.03</td>
</tr>
<tr>
<td>Cumulus PD (%)</td>
<td>27.44 (15.10)</td>
<td>25.35 (15.12)</td>
<td>26.09 (15.13)</td>
<td>0.22</td>
</tr>
<tr>
<td>PXA Dense Volume (cm³)</td>
<td>95.32 (60.13)</td>
<td>76.14 (51.05)</td>
<td>82.93 (55.13)</td>
<td>0.01</td>
</tr>
<tr>
<td>PXA VPD (%)</td>
<td>14.39 (8.92)</td>
<td>12.41 (6.80)</td>
<td>13.11 (7.66)</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 2. OR and AUC values for Cumulus and PXA. * Adjusted by age, BMI, race and menopausal status.

<table>
<thead>
<tr>
<th>Measurement Quartile</th>
<th>No. Cases</th>
<th>No. Controls</th>
<th>Quart. unadjust. OR (95% CI)</th>
<th>Quart. adjust. OR (95% CI)</th>
<th>SD unadjust. AUC</th>
<th>SD adjust. AUC</th>
<th>AUC (95% CI)</th>
<th>Adjust. AUC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulus Dense Area</td>
<td>131</td>
<td>239</td>
<td>1.31 (1.05,1.71)</td>
<td>1.38 (1.08,1.76)</td>
<td>0.56</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1 (0.96)</td>
<td>16</td>
<td>60</td>
<td>1.00 (Ref.)</td>
<td>1.00 (Ref.)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q2 (14.23)</td>
<td>43</td>
<td>59</td>
<td>2.73 (1.39,5.38)</td>
<td>3.7 (1.70,8.07)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3 (27.61)</td>
<td>32</td>
<td>60</td>
<td>2.09 (0.40,4.02)</td>
<td>2.71 (1.16,6.32)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q4 (40.12)</td>
<td>40</td>
<td>60</td>
<td>2.5 (1.26,4.94)</td>
<td>2.84 (1.20,7.63)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulus PD</td>
<td>131</td>
<td>239</td>
<td>1.15 (0.92,1.43)</td>
<td>1.40 (1.04,1.90)</td>
<td>0.34</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1 (0.59)</td>
<td>26</td>
<td>60</td>
<td>1.00 (Ref.)</td>
<td>1.00 (Ref.)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q2 (12.04)</td>
<td>30</td>
<td>59</td>
<td>1.17 (0.62,2.22)</td>
<td>1.54 (0.74,3.21)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3 (25.96)</td>
<td>39</td>
<td>60</td>
<td>1.5 (0.81,2.8)</td>
<td>2.24 (0.95,5.27)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q4 (36.36)</td>
<td>36</td>
<td>60</td>
<td>1.38 (0.75,2.57)</td>
<td>2.86 (1.08,7.58)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PXA Dense Volume</td>
<td>131</td>
<td>239</td>
<td>1.41 (1.13,1.75)</td>
<td>1.43 (1.11,1.81)</td>
<td>0.50</td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1 (4.10)</td>
<td>18</td>
<td>60</td>
<td>1.00 (Ref.)</td>
<td>1.00 (Ref.)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q2 (42.42)</td>
<td>32</td>
<td>59</td>
<td>1.81 (0.92,3.57)</td>
<td>1.75 (0.84,3.65)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3 (53.88)</td>
<td>32</td>
<td>60</td>
<td>1.78 (0.90,3.51)</td>
<td>1.82 (0.83,4.00)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q4 (97.23)</td>
<td>49</td>
<td>60</td>
<td>2.72 (1.42,5.20)</td>
<td>3.26 (1.47,7.20)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PXA VPD</td>
<td>131</td>
<td>239</td>
<td>1.29 (1.04,1.59)</td>
<td>1.32 (1.17,2.00)</td>
<td>0.56</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1 (4.07)</td>
<td>24</td>
<td>60</td>
<td>1.00 (Ref.)</td>
<td>1.00 (Ref.)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q2 (7.45)</td>
<td>35</td>
<td>59</td>
<td>1.48 (0.79,2.79)</td>
<td>1.98 (0.97,4.04)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3 (10.54)</td>
<td>27</td>
<td>60</td>
<td>1.12 (0.58,2.17)</td>
<td>1.32 (0.60,2.92)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q4 (15.47)</td>
<td>45</td>
<td>60</td>
<td>1.88 (1.02,3.45)</td>
<td>5.30 (0.99,14.16)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References

Characterizing the Frequency of Pharmacogenomic Biomarker-Guided Prescribing for Drugs with Pharmacogenomic Biomarker information in the FDA Labelling: A Pilot Study Using Data from an Electronic Health Record

Beth Devine, PhD, PharmD, MBA1,2,3, Maher Khelifi, PharmD2, Katelyn Keyloun, PharmD1, Nathaniel Hendrix, PharmD1, Patrick Mathias, MD, PhD4, Christian Bock1,4+, Peter Tarczy-Hornoch, MD, FACMI1,4,6

1Pharmaceutical Outcomes Research and Policy Program; 2Department of Biomedical Informatics and Medical Education; 3Department of Health Services; 4Department of Laboratory Medicine, 5Department of Pediatrics, Division of Neonatology; 6Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

*(now University of Heidelberg, Germany)

BACKGROUND
Based on self-report, approximately 39% of United States-based physicians obtain pharmacogenomic (PGx) testing information from drug labelling. However, physicians may broadly underutilize this information to guide practice.1 The recent large-scale availability of clinical data from electronic health records (EHRs), coupled with the rapid decrease in the per-test cost of genotyping, and the capability of advanced clinical decision support systems to guide therapy, make the investigation of PGx-guided prescribing in real-world settings an area of research that is now feasible and timely. The objective of this pilot study is to use clinical data from the EHRs and other clinical systems within the University of Washington Medicine system (inpatient and outpatient), to estimate and characterize PGx biomarker use, assess the association between characteristics of PGx-med pairs and the use of PGx tests, and estimate the temporal association between medication prescribing and PGx testing.

METHODS
Using our institutional clinical data warehouse, we extracted data from 2 EHRs and an organization-wide laboratory information system from January 2010 through March 2014. We identified patients ≥ 18 years of age who had an incident, index prescription (for outpatients) or medication administration event (for inpatients) for the medications which contain a drug label the includes a recommendation for PGx testing. By linking laboratory data with index medication events, we estimated the frequency of associated PGx-biomarker testing for this group of medications, and established the temporal association between each medication order and PGx testing order. In addition, we examined the medications with the highest proportion of PGx testing to examine differences by therapeutic area.

RESULTS
The patient cohort consists of 188,783 patients who each received one or more of 137 medications that correspond with one, one or more of 47 distinct PGx biomarkers, which together constitute 155 PGx-med pairs. The dataset captures 408,603 index prescriptions (outpatient) or medication administration events (inpatient) and 22,441 PGx tests. Across all eligible index prescriptions in the study period, 0.9% were associated with a PGx test, with an increase from 0.77% of index medications in 2010 to 1.12% in 2013 (Figure 1). Greater than 50% of PGx tests were performed within 60 days of the index medication prescription/administration (Figure 2). The top 25 most frequently prescribed/administered medications accounted for 81% of index medications but only 0.3% of such events had an associated PGx test performed. Oncology medications accounted for 17 of the top 20 medications (85%) with the highest rate of associated PGx testing (Table 1).

DISCUSSION
By integrating data from multiple EHRs and other clinical systems, we examined over 400,000 eligible medication prescriptions and administration events. We determined that <1% had a corresponding PGx biomarker test, and the low overall percentage was driven by frequently ordered medications such as omeprazole or simvastatin that rarely had an associated PGx test performed. Over the 4 full years covered by the study, the use of PGx testing has increased with time. A majority of the PGx tests we captured were performed within 2 months of the index medication event. As expected, oncology medications represented a significant majority of the top 20 medications with the highest proportion of associated PGx testing. Our study informs the field of PGx biomarker use by providing an estimate of the frequency of use in real-world clinical practice, enabling investigation of the characteristics of PGx-med pairs most frequently associated with biomarker use, and estimating the temporal association between incident medication use and biomarker testing.


30
Figure 1. Cumulative percentage of medication events with associated PGx test over time.

Figure 2. Distribution of durations between index medication event and PGx test order.

<table>
<thead>
<tr>
<th>Medication</th>
<th>Medication Type</th>
<th>Number of Patients</th>
<th>Patients With PGx Test</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRENXTUMAB</td>
<td>Oncology</td>
<td>22</td>
<td>21</td>
<td>96%</td>
</tr>
<tr>
<td>TOSITUMOMAB</td>
<td>Oncology</td>
<td>36</td>
<td>32</td>
<td>89%</td>
</tr>
<tr>
<td>ARSENIC</td>
<td>Oncology</td>
<td>20</td>
<td>16</td>
<td>80%</td>
</tr>
<tr>
<td>VEMURAFENIB</td>
<td>Oncology</td>
<td>15</td>
<td>12</td>
<td>80%</td>
</tr>
<tr>
<td>OFATUMUMAB</td>
<td>Oncology</td>
<td>33</td>
<td>26</td>
<td>79%</td>
</tr>
<tr>
<td>PERTUZUMAB</td>
<td>Oncology</td>
<td>58</td>
<td>38</td>
<td>66%</td>
</tr>
<tr>
<td>TRASTUZUMAB</td>
<td>Oncology</td>
<td>409</td>
<td>256</td>
<td>63%</td>
</tr>
<tr>
<td>DAPSONE</td>
<td>Infectious Diseases Dermatology</td>
<td>1095</td>
<td>665</td>
<td>61%</td>
</tr>
<tr>
<td>ADO-TRASTUZUMAB</td>
<td>Oncology</td>
<td>25</td>
<td>14</td>
<td>56%</td>
</tr>
<tr>
<td>PRIMAQUINE</td>
<td>Infectious Diseases</td>
<td>44</td>
<td>24</td>
<td>55%</td>
</tr>
<tr>
<td>RITUXIMAB</td>
<td>Oncology</td>
<td>1402</td>
<td>764</td>
<td>55%</td>
</tr>
<tr>
<td>PANITUMUMAB</td>
<td>Oncology</td>
<td>38</td>
<td>19</td>
<td>50%</td>
</tr>
<tr>
<td>LENALIDOMIDE</td>
<td>Hematology</td>
<td>142</td>
<td>69</td>
<td>49%</td>
</tr>
<tr>
<td>FULVRENTANT</td>
<td>Oncology</td>
<td>143</td>
<td>67</td>
<td>47%</td>
</tr>
<tr>
<td>LAPATINIB</td>
<td>Oncology</td>
<td>13</td>
<td>6</td>
<td>46%</td>
</tr>
<tr>
<td>NILOTINIB</td>
<td>Oncology</td>
<td>21</td>
<td>8</td>
<td>38%</td>
</tr>
<tr>
<td>RASBSURICASE</td>
<td>Oncology</td>
<td>106</td>
<td>40</td>
<td>38%</td>
</tr>
<tr>
<td>DASSTATINIB</td>
<td>Oncology</td>
<td>57</td>
<td>19</td>
<td>33%</td>
</tr>
<tr>
<td>TRAMETINIB</td>
<td>Oncology</td>
<td>3</td>
<td>1</td>
<td>33%</td>
</tr>
<tr>
<td>IMATINIB</td>
<td>Oncology</td>
<td>87</td>
<td>27</td>
<td>31%</td>
</tr>
</tbody>
</table>

Table 1. Top 20 medications by percentage of index medication events with an associated PGx test.
2 Years Later: Follow-up to Analysis of Electronic Medication Orders with Large Overdoses

Judith W Dexheimer, PhD, Eric S Kirkendall, MD, MS, Michal Kouril, PhD, Thomas Minich, BSpH, Philip Hagedorn, MD, Cecilia Mahdi, MBA, S. Andrew Spooner MD
Cincinnati Children’s Hospital Medical Center (CCHMC), Cincinnati, OH

Introduction

A major challenge facing pediatric medicine is the management of weight-based dosing rules in electronic health records (EHRs). EHRs implement rules supplied by third-party vendors to help guide the dosing process. Since many of these rules are conservative, they result in noisy alerting, and are therefore overridden by users. EHR implementers typically customize these vendor-supplied rules to reduce this noise. In a previous report, we examined users’ responses to large overdoses and characterized user-CPOE actions associated with these large overdoses. In this paper, we provide an update to our original manuscript with 24 months’ additional data. We track the original trends and how well our medication rules are performing.

Methods

Cincinnati Children’s Hospital Medical Center (CCHMC) is a 587-bed, freestanding, academic, pediatric medical center with more than 1.2 million patient encounters annually and over 800 faculty members. It has over 30,000 admissions, 33,000 surgeries, 900,000 ambulatory encounters, and 125,000 emergency department visits per year. CCHMC uses an enterprise EHR system from Epic™ (Verona, WI), initially configured to use the Medi-Span database that includes dosing rules (Wolters Kluwer Health, Philadelphia, PA). The CCHMC pharmacy team customizes these rules per local dosing guidance. Up to 80% of alerts are triggered by customized rules (Figure 1). There are approximately 28,684 medication orders written monthly; approximately 87% of the alerts related to these medications are overridden at the time of ordering (Figure 2). Of these, approximately 5,831 are non-filtered single-dose overdose alerts seen at the order-entry phase; 90% of these are overridden. We focused on large (≥500% of the maximum rule-based dose) overdoses. We assume that if a user cancels the ordering session after seeing the alert dialog box, it is likely that the user cancelled the orders because of one of the presented alerts, so we define alert salience as the number of orders associated with an alert cancellation divided by the total number of alerts presented.

Results and Discussion

As the result of regular modification of dosing rules over several years, we have seen a decrease in the number of alerts shown to users. The updated data are shown in figures 2-4. We have also seen a decrease in % of all orders that had a single dose overdose alert by 3% from a peak of around 10% to roughly 7% (Figure 1). With this decrease in large overdose alerts (Figure 3), we have seen an increase in alert salience (Figure 4). An example of the increased salience of alerts as a result of rule modification is illustrated in figure 4 for clindamycin. The user burden
for alerts in our system is small (8% of orders) (Figure 3). Over the course of tailoring alert parameters, we have seen this burden fall from 10% to 7%. We have evidence to suggest that decreasing this alert burden increases salience.

**Figure 2.** Medications with >= 500% overdose alerts.

**Figure 3.** Order Alert Rates in enter orders.

**Figure 4.** Salience rates for single dose over dose alerts and for medications with >=500% overdose rates.

**References**
Completeness and Timeliness of Notifiable Disease Surveillance Data 
Submitted by Providers to Public Health Authorities 

Brian E. Dixon1,2, Patrick T.S. Lai3, Uzay Kirbiyik1, Zuoyi Zhang2, Debra Revere4, Rebecca Hills4, P. Joseph Gibson5, Jennifer Williams2, Shaun J. Grannis2
1IU Richard M. Fairbanks School of Public Health, Indianapolis, IN; 2Regenstrief Institute Center for Biomedical Informatics, Indianapolis, IN; 3IU School of Informatics and Computing, Indianapolis, IN; 4University of Washington, Seattle, WA; 5Marion County Public Health Department, Indianapolis, IN

Introduction
Surveillance is the cornerstone of public health practice. Traditionally, health departments have relied on passive reporting of communicable and infectious diseases by hospital, laboratory and clinic staff to public health authorities. However, passive approaches are known to be burdensome for reporters, producing incomplete and delayed reports which can impede assessment and potentially delay recognition of patterns and outbreaks.

Increasingly, surveillance practice is shifting toward greater use of electronic receipt of disease information. The adoption of electronic health record (EHR) systems and health information exchange (HIE) among clinical organizations, driven by policies like the ‘meaningful use’ program, is creating an information infrastructure that public health organizations can leverage. Data from the U.S. Office of the National Coordinator for Health Information Technology indicate that health departments receive up to 62% of their total laboratory-based reports for notifiable diseases by electronic transmission (1). Although provider-based reporting continues to be paper-based, EHR systems and HIE networks provide a framework for electronic submission of treatment, corollary results, and other details not available from laboratory information systems.

Methods
We are conducting a controlled trial of an informatics intervention designed to improve the completeness and timeliness of provider-based information to public health authorities for notifiable disease cases. Our study occurs in the context of the Indiana Health Information Exchange (IHIE), a large, robust HIE network that delivers clinical results to over 25,000 physicians. Using components within the IHIE information infrastructure, including the Notifiable Condition Detector, our intervention is pre-populating the official Indiana State Department of Health communicable disease reporting form (paper) with patient demographics, notifiable disease confirmatory test results, and provider information using ELR data. The pre-populated form is delivered electronically to the provider using the HIE for review and completion. A process that previously began with a human in the clinic noticing a result which should be reported to public health then completing a blank, paper form that was eventually faxed to the local health department. A detailed description of the study protocol was described previously (2).

The study focuses on seven representative diseases (Salmonella, Hepatitis C, Hepatitis B, Chlamydia, Gonorrhea, Syphilis, Histoplasmosis) commonly investigated by local health department staff. Like most states, Indiana utilizes a dual-reporting structure: mandatory case reporting by providers and reporting of test results by laboratories. Prior to deploying the intervention, we gathered baseline reporting information from fax and paper notifiable disease reports submitted by both providers and labs to the Marion County Public Health Department. The time period for collecting data varied based on disease prevalence. We gathered reports for more prevalent diseases (e.g., Chlamydia) over a three-month period (May to July 2012), while reports for less prevalent diseases (e.g., Histoplasmosis) were gathered over two years (2010-2012). We similarly gathered electronic laboratory reports for the same conditions within the same timeframe from the NCD within the IHIE infrastructure.

Key information, that which is used by the health department to initiate case investigation, was abstracted from each report and collated by disease. We evaluated the completeness of reporting data elements for paper, fax-based, and electronic lab reports sent from providers and labs using methods previously described in prior work on just electronic lab reports (3, 4). Timeliness (elapsed time) was measured using the difference in number of calendar days between the date of the positive lab result and the date which the report was submitted to the first public health authority (local health department or state health agency).

Results
A total of 12,098 reports representing 8,353 unique patients provided the baseline dataset. Providers submitted 2,740 paper/fax reports for 2,314 patients; labs submitted 1,447 paper/fax reports for 1,134 patients; and labs submitted
7,911 ELRs for 6,294 patients. Some patient files (N=2,450) included multiple, duplicate reports from providers and/or labs. For example, in some cases both the infection control staff at a hospital and the patient’s primary care provider submitted a report. Table 1 describes the completeness of data fields on paper, fax-based and ELR reports of notifiable diseases.

Table 1 – Comparison of data element completeness for provider-based paper reports, lab-based fax reports and electronic lab-based reports submitted to public health for seven notifiable conditions.

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Provider Reports (Paper) N=2740</th>
<th>Lab Reports (Fax) N=1447</th>
<th>Electronic Lab Reports N=7911</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>count</td>
<td>%</td>
<td>count</td>
</tr>
<tr>
<td>Patient Name</td>
<td>2740</td>
<td>100</td>
<td>1447</td>
</tr>
<tr>
<td>Patient Address</td>
<td>1644</td>
<td>60</td>
<td>850</td>
</tr>
<tr>
<td>Zip Code</td>
<td>1584</td>
<td>57.8</td>
<td>853</td>
</tr>
<tr>
<td>Phone Number</td>
<td>1498</td>
<td>54.7</td>
<td>901</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>2661</td>
<td>97.1</td>
<td>1381</td>
</tr>
<tr>
<td>Sex</td>
<td>2481</td>
<td>90.5</td>
<td>1415</td>
</tr>
<tr>
<td>Race</td>
<td>1495</td>
<td>54.6</td>
<td>207</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>1031</td>
<td>37.6</td>
<td>206</td>
</tr>
<tr>
<td>Physician First Name</td>
<td>1586</td>
<td>57.9</td>
<td>1205</td>
</tr>
<tr>
<td>Physician Last Name</td>
<td>1632</td>
<td>59.6</td>
<td>1272</td>
</tr>
<tr>
<td>Physician Address</td>
<td>1275</td>
<td>46.5</td>
<td>1290</td>
</tr>
<tr>
<td>Physician Zip Code</td>
<td>921</td>
<td>33.6</td>
<td>1273</td>
</tr>
<tr>
<td>Physician Phone</td>
<td>1240</td>
<td>45.3</td>
<td>1148</td>
</tr>
<tr>
<td>Lab Test Name</td>
<td>2081</td>
<td>75.9</td>
<td>1439</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Mean</td>
<td>Range</td>
<td>Mean</td>
</tr>
<tr>
<td>No. Days</td>
<td>9.74</td>
<td>(0-375)</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Discussion

Completeness of data in notifiable disease reports is heterogeneous. In general ELR data are more complete than paper- and fax-based reports from labs and providers, with the exception of provider contact information which is more complete in faxed lab reports and ethnicity which is more complete in provider reports. Interestingly, lab-based reports often contained more complete provider contact data than documents submitted from clinics and infection control. With respect to timeliness, the mean difference in days for ELR is superior to both faxed-based lab and provider reports. The results are similar to our prior analyses involving ELR data sources (3-5).

Completeness patterns call into question the need for dual-reporting laws in states that require both providers and labs to report the same information to public health for accurate surveillance of notifiable diseases. Laboratory data sources possess unique strengths with only minor weaknesses. Given that EHR and HIE systems are increasingly interfaced with health departments for ELR to achieve the aims of the meaningful use policies, there is an opportunity to supplant manual, paper-based processes with enhanced data from integrated electronic data sources. This could be a victory not only for health departments who desire more complete and timely data but also providers who find public health reporting processes time-consuming and challenging given competing priorities.

References


Pharmacy drug dispensing after physician discontinuation (cancel) orders

Tewodros Eguale, MD, PhD1,2,3,4, Aman Verma, PhD4, Enrique Seoane-Vazquez, PhD1,2, Rosa Rodriguez-Monguio, PhD5, David W. Bates, MD, MSc2,3, Robyn Tamblyn, PhD4 and Gordon D. Schiff, MD2,3

1. Pharmacoeconomics and Health Policy, MCPHS University, Boston, MA, United States;
2. Division of General Medicine and Primary Care, Brigham and Women Hospital, Boston, MA, United States;
3. Harvard School of Medicine, Boston, MA, United States and
4. Clinical and Health Informatics Research Group, McGill University, Montreal, Quebec, Canada.
5. School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA, USA

Background: Physicians discontinue previously prescribed drugs due to considerations of safety, effectiveness, and cost, but it is unclear what effect these discontinuation orders have on pharmacy dispensing. New electronic health records (EHR) allow us to monitor discontinuation orders, discontinuation reasons, and actual dispensation data.

Objectives: Using data from an EHR, we assessed the frequency of pharmacy drug dispensation after physicians discontinue orders, and modeled how the discontinuation reasons affected pharmacy dispensation.

Methods: We conducted the study in two Canadian cities where family physicians use the MOXXI EHR, which allows physicians to prescribe and discontinue drugs and send the information to the pharmacy. We included all discontinuation orders issued between 2005 and 2012 for patients fully covered by provincial drug insurance. Using logistic regression, we modeled pharmacy dispensation of discontinued drugs within 12 months of the discontinuation order as a function of the discontinuation reasons, and patient and drug characteristics.

Results: Between 2005 and 2012 there were 40,452 drug discontinuation orders, of which 7,325 (18.1%) were dispensed within 12 months (75% within the 1st month). Many drugs were discontinued because they were “no longer necessary” (27.3%), ineffective (18.7%) or caused adverse drug events (18.2%). Drugs discontinued due to adverse drug event [OR: 0.7 95% CI: (0.70.8)] and allergic reaction [OR: 0.6 95% CI: (0.41.0)] had low risk of dispensation compared to drugs stopped because they were “no longer necessary”, while drugs discontinued because they were ineffective had a higher odds [OR: 1.3 95% CI: (1.21.5)]. Compared to gastrointestinal drugs, central nervous system drugs [OR: 1.5 95% CI: (1.31.6)] and cardiovascular drugs [OR: 1.1 95% CI: (1.01.2)] had higher odds of dispensation and anti-infective drugs had lower odds [OR: 95% CI: 0.5 (0.40.6)].

Conclusions: Patient safety may have been compromised when discontinued drugs were dispensed to patients, especially those discontinued due to an adverse drug reaction. The next generation of electronic health records should better integrate with the pharmacy, to avert dispensation of discontinued drugs and curb cost to the patient and the payer.
Computer-Mediated Intervention to Improve Medication Literacy in Seniors with Diabetes Results in Better Patient-Reported Outcomes and Glycemic Control

Joseph Finkelstein, MD, PhD, McKenzie E. Bedra, MPH
Chronic Disease Informatics Program, Johns Hopkins University, Baltimore, MD

Abstract

The purpose of this study was to compare interactive continuous patient education (iCOPE) to a brochure in a randomized controlled trial. Older adults taking oral diabetes medications were randomly assigned to receive the same diabetes medications curriculum via iCOPE or printed brochure combined with periodic reminders. At 3 months significant improvement in medication knowledge, self-efficacy and glycemic control was found in the intervention group whereas changes in the control group remained not significant.

Introduction

Low medication literacy was shown to be prevalent in older adults and to negatively affect quality of life and clinical outcomes. Ongoing patient empowerment addressing medication literacy results in patient activation and improved health outcomes. Interactive health communication technologies have been shown effective in promoting health education and improving medication literacy, however, its potential in older adults has not been uncovered systematically.

Methods

Interactive continuous patient education (iCOPE) system was used to deliver medication literacy intervention via touch screen tablet computers using personalized avatar-based interface. The platform supports continuous patient engagement and empowerment driven by adult learning theories as previously described [1]. Eligible participants included older adults who were 65 years of age and older, currently taking or have taken an oral diabetes medication in the past, and had a Mini Mental State Examination score of at least 23 [1]. The study subjects were randomly assigned to receive the same curriculum via iCOPE or a printed brochure. The enrolled participants were asked to complete the oral diabetes medication curriculum within 24 hours at their own pace. During 6-month follow-up, patients in the intervention group periodically received interactive engagement materials via e-mail or text messages whereas the patients in the control group received printed materials via mail on a regular basis. The participants were assessed at baseline, 24-hr, 1-mm, 3-mm, and 6-mm visits using established questionnaires reflecting medication literacy, self-efficacy, quality of life and self-management skills. In addition, HbA1c was measured at each visit but 24-hr. Socio-demographic background, disease history, health and computer literacy were collected at baseline.

Results

Overall, 101 older adults were randomized to intervention and control groups with no differences at baseline; 59% were females and the average age was 75; 81% responded that the tablet education was not difficult to use, whereas in the control group 59% reported the sentences used in the educational materials were difficult. At 3 months significant improvement in medication knowledge, self-efficacy and glycemic control was found in the intervention group whereas changes in the control group remained not significant (Table 1). There was a threefold difference in mean knowledge score increase (6.5±5.8 vs. 2.0±3.4) and 4-fold difference in mean HbA1c decrease (-0.41±1.24 vs. -0.13±1.48) between intervention and control groups at 3 months compared to baseline (Figure 1).

Discussion and Conclusion

Education via a touch screen tablet resulted in knowledge gains. Continuous patient engagement and empowerment delivered via interactive mobile tools is effective means to improve medication literacy in older adults with diabetes.

References

Table 1. Comparison of intervention and control groups at baseline and 3-month follow-up visits [mean(SD)].

<table>
<thead>
<tr>
<th>Medication Knowledge Test Score</th>
<th>Baseline (BL)</th>
<th>24 Hrs</th>
<th>1 Month</th>
<th>3 month</th>
<th>p-value(BL vs.3Mon)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control</strong></td>
<td>28.25(4.35)</td>
<td>30.54(5.23)</td>
<td>30.68(5.56)</td>
<td>30.39(6.35)</td>
<td>0.0021*</td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>26.59(5.73)</td>
<td>33.62(3.90)</td>
<td>32.82(3.91)</td>
<td>33.10(3.31)</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td><strong>Patient Activation Measure</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>85.98(8.73)</td>
<td>87.93(7.33)</td>
<td>88.40(7.76)</td>
<td>87.34(10.18)</td>
<td>0.5246</td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>88.56(12.19)</td>
<td>92.10(9.68)</td>
<td>92.96(9.85)</td>
<td>91.85(11.73)</td>
<td>0.0315*</td>
</tr>
<tr>
<td><strong>Medication Communication Self-Efficacy (AURA)</strong></td>
<td>Baseline</td>
<td>24 Hrs</td>
<td>1 Month</td>
<td>3 month</td>
<td>p-value(BL vs.3Mon)</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>15.23(1.21)</td>
<td>15.54(1.39)</td>
<td>15.58(0.70)</td>
<td>15.54(0.90)</td>
<td>0.3257</td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>15.10(1.56)</td>
<td>15.10(2.06)</td>
<td>15.20(1.77)</td>
<td>15.77(0.50)</td>
<td>0.0254*</td>
</tr>
<tr>
<td><strong>Medication Understanding Self-Efficacy (MUSE)</strong></td>
<td>Baseline</td>
<td>24 Hrs</td>
<td>1 Month</td>
<td>3 month</td>
<td>p-value(BL vs.3Mon)</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>29.92(3.75)</td>
<td>31.00(1.88)</td>
<td>30.62(2.48)</td>
<td>30.23(4.05)</td>
<td>0.3954</td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>29.28(5.44)</td>
<td>31.14(1.53)</td>
<td>29.97(4.82)</td>
<td>30.69(4.49)</td>
<td>0.0395*</td>
</tr>
<tr>
<td><strong>Medication Adherence Self-Efficacy (MASES)</strong></td>
<td>Baseline</td>
<td>24 Hrs</td>
<td>1 Month</td>
<td>3 month</td>
<td>p-value(BL vs.3Mon)</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>2.79(0.28)</td>
<td>2.86(0.18)</td>
<td>2.82(0.21)</td>
<td>2.87(0.22)</td>
<td>0.1192</td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>2.84(0.23)</td>
<td>2.89(0.18)</td>
<td>2.88(0.23)</td>
<td>2.92(0.14)</td>
<td>0.0031*</td>
</tr>
<tr>
<td><strong>Medication Adherence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>8.70(2.40)</td>
<td>9.22(2.22)</td>
<td>9.48(2.34)</td>
<td>0.0471*</td>
<td></td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>8.95(1.76)</td>
<td>9.67(1.46)</td>
<td>10.03(1.25)</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td><strong>HbA1C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>7.31(1.54)</td>
<td>6.97(1.48)</td>
<td>7.15(1.46)</td>
<td>0.8049</td>
<td></td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>7.56(1.55)</td>
<td>7.28(1.26)</td>
<td>7.09(1.35)</td>
<td>0.0163*</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Mean outcome change over time.
Feature Selection Based LapSVM to Classify Medical Event Reports and Enhance Patient Safety

Samah Jamal Fodeh, PhD¹, Cynthia Brandt, MD MPH², Perry Miller, MD, PhD³, Michele Koss RN⁴, Andrea Benin, MD⁵

(¹-⁴) Yale University School of Medicine, Yale University, New Haven, CT, 06520
(⁴,⁵) Connecticut Children’s Medical Center, Hartford, CT

Introduction. Timely reporting and analysis of adverse events and medical errors is critical to driving forward programs in patient safety. Due to the large numbers of event reports accumulating daily in health institutions, manually finding and labeling certain types of events needs to be replaced with automated classification methods. In semi-supervised learning, methods usually utilize unlabeled data in addition to labeled data to perform a certain classification task. In this work we propose feature based Laplacian support vector machine (FS-LapSVM) to identify two major error events: patient mismatches errors and patient weight errors.

Problem. Event reporting systems rapidly accumulate incident reports, therefore the manual review of the reports became very challenging and prohibitively resource-intensive. Although most reporting systems provide a coding system for classifying major patient safety events, labeling a report using the coding system remains a subjective matter determined by the person documenting the report which results in inconsistency in how reports are classified. Also, coding systems do not necessarily provide a match to the event to be classified. One alternative to address these issues is to automatically classify event reports by exploiting their narratives using text-based classification from machine learning. While classification has been successfully used to detect AE in the EHRs, its utility has been relatively unexplored on event and medical error reports.

Methods In semi-supervised learning, a classification function is estimated using labeled and unlabeled data. We propose FS-LapSVM; a hybrid method that combines F-score feature selection [1] with Laplacian support vector machine [2]. Although many feature selection approaches have been combined with LapSVM, the F-score selection approach has not been tested before with LapSVM. We labeled (~9000) reports that we extracted from Quintros reporting system at Connecticut Children’s Medical Center. The labeled reference standard contained 1102, 85 reports of patient mismatches and weight errors respectively and the remaining reports were assigned the category “other”. Because LapSVM utilizes labeled and unlabeled reports, for each event category, the sample included three types of reports: (1) labeled reports from the category to be detected, (2) equivalent number of records from the “other” category and (3) unlabeled reports equivalent in size to 15% of both categories.

Results: With 20% reduction of features, Table 1 shows that FS-LapSVM classifier outperformed standard LapSVM in classifying patient mismatch across all metrics on the validation set. Superior performance of FS-LapSVM finding weight error reports was observed compared to LapSVM; an increase of 10%, 5%, 5% in terms of PPV, F-measure, and accuracy, respectively when the respective models were evaluated on the validation set.

Table 1: Performance comparison between FS-LapSVM and LapSVM

<table>
<thead>
<tr>
<th></th>
<th>Patient Mismatch Event</th>
<th>Weight Mismatch Event</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LapSVM</td>
<td>FS-LapSVM</td>
</tr>
<tr>
<td>PPV</td>
<td>0.8</td>
<td>0.82</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.78</td>
<td>0.79</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.81</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Conclusions. We demonstrate the utility of FS-LapSVM which we utilized to identify two important events: patients’ mismatches and weight errors. FS-LapSVM can effectively decrease the size of the feature set and achieve better results compared to LapSVM. Next, we will explore the effect of adding more unlabeled data on performance.

References
Developing a Collaborative Evaluation Framework for Utah’s APCD

Jennifer H. Garvin, PhD, MBA, RHIA1, 2, 3, Kristina Doing-Harris, PhD1, Jonathan H. Cardwell, MS1, Kailah Davis, PhD4, Charles Hawley, MA4, and Wu Xu, PhD1, 2, 4, 1Department of Biomedical Informatics, 2Division of Epidemiology, University of Utah, Salt Lake City, UT; 3VA Health Care System, Salt Lake City, UT; 4Utah Department of Health

Abstract: We developed an evaluation framework based on a collaborative evaluation model for a large-scale, federally-funded information system improvement project with the Utah Department of Health. In this presentation, we describe the development of the framework and its elements, all of which form an essential foundation with regard to the objectives of Utah’s All Payer Claims Database (APCD) as well as process evaluation results.

Introduction: The overall goal for the Cycle III grant is to improve the Utah All Payers Claims Database (APCD)1 by developing analytic capacity and producing online pricing/cost transparency reports for consumers, employers, researchers, and the general public in Utah. In order to facilitate success for this large-scale APCD implementation, we chose a collaborative evaluation model that paralleled the project management process chosen by the APCD principal investigator2. We present the results of the evaluation at year which relates to processes used. Evaluation of use cases will occur in year 3. Prior APCD development efforts did not include collaborative evaluation methods.

Methods: Development of the evaluation framework began by identifying our process of change, which brings the entire evaluation program to a concise focus. Once identified, evaluation questions were formed to support that process of change, delineating what, when, who, and how the data will be collected until each evaluation question has been answered. Logic models are also developed collaboratively with project groups. This collaborative process between the project groups and evaluation team set a foundation from which to monitor and evaluate the attainment of the project’s objectives and facilitated the use of evaluation data by the project teams. Evaluation components were shared and iteratively developed with project groups via email and meeting presentations.

Results: Through a total of 71 communications with 16 project groups, we determined our evaluation plan. During this iterative development we also determined key use cases to drive improved APCD data quality, new health information technology development, and analytic capability expansion using a total of 274 communications. The ability to undertake uses cases will be assessed in year 3. During year 1 and 2 we evaluated the use of components of best practices for improving claims data quality, large-scale health IT (HIT) development, healthcare information security and privacy, and software development. As reflected in Table 1, there was significant and increasing use of component parts of best practices. We also used logic models with each team.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Year 1</th>
<th></th>
<th>Year 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Partial Use</td>
<td>Ongoing Use</td>
<td>Complete Use</td>
<td>Ongoing Use</td>
</tr>
<tr>
<td>Data Quality</td>
<td>0</td>
<td>7</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>HIT Development</td>
<td>4</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Privacy and Security</td>
<td>1</td>
<td>4</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>Software Development</td>
<td>4</td>
<td>20</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>Totals:</td>
<td>9</td>
<td>43</td>
<td>54</td>
<td>46</td>
</tr>
</tbody>
</table>

Conclusion: Our collaborative evaluation framework was instrumental in the development and success of Utah’s APCD implementation to date. The framework elements, such as best practices, are foundational to the support of the use cases and are essential to accomplishing the project’s objectives. Collaborative program evaluation processes provided additional data, processes, and value to facilitate successful completion of the project. The evaluation process will benefit stakeholders by improving online pricing/cost transparency reports for consumers, employers, researchers, and the general public in Utah.

Acknowledgements: We thank the stakeholders and members of the Utah Cycle III project team for their engagement with the program evaluation team. This publication is funded by CMS Grants to States to Support Health Insurance Rate Review and Increase Transparency in Health Care Pricing, Cycle III [Grant Number: 1 PRPPR140059-01-00], through a subcontract with Utah Department of Health Office of Health Care Statistics.

References

Sheila M. Gephart, PhD, RN¹, Joanna Abraham, PhD,², Rebecca Raszewski,², Lauren Price, BSN, RN,¹,³ Karen Dunn-Lopez, MPH, PhD, RN²

¹The University of Arizona, Tucson, AZ; ²The University of Illinois at Chicago, Chicago, Ill.; ³Southern Arizona Veterans Affairs

**Introduction**

Clinical decision support (CDS) is becoming an increasingly important component of health care designed to improve the clinical processes and outcomes of care (1). While much has been written about CDS targeting physicians’ decision making (2, 3), comparatively less is known about the outcomes for CDS targeting nurses, particularly for bedside acute care nurses (4, 5). In this abstract, we report on our preliminary results of an integrative review of studies regarding CDS for acute care bedside nurses.

**Methods**

We conducted a comprehensive integrative review (6) of published research articles on the design and evaluation of nursing CDS. To avoid overlap with previous reviews, we included articles published from 2006-2013. Our search strategy included original studies identified across 6 databases (CINAHL, EMBASE, IEEE Explorer, MEDLINE via PubMed, Scopus, and Web of Science). We used the following subject headings and key words: computer-assisted decision making, computers, clinical decision support systems, expert systems, information systems, hospital, nursing informatics, nursing homes, nursing staff, reminder systems, and derivatives of the word, “nurse”. Inclusion criteria included: peer-reviewed articles, focused on nursing decision support in an acute care settings. Exclusion criteria included: conference proceedings, articles without exclusive focus on RN (Registered Nurse)-directed decisions, CDS that was not computerized, non-acute settings, or included RN’s in combination with other clinicians. A structured data form and definitions including purpose, methods, study design and outcomes, and integration with Electronic Health Record (EHR) was developed for data abstraction and used for data analysis.

**Results**

Our search yielded 2477 unique articles that were systematically reduced to 28 articles that met the inclusion criteria (see Figure 1). Inter-rater agreement was established at 89-91% across the steps of review and data abstraction. Of 28 articles included, randomized controlled trials (n=1), quasi-experimental (n= 11), single correlational (n= 4) and single descriptive or qualitative (n=12) were identified. The purposes of CDS in acute care nursing was most often to enforce guideline adherence (36%), and support nursing diagnostic decision making (36%). Less often, CDS was used to support medication dosing with algorithms, situational awareness and emergency triaging. CDS was integrated into the EHR less than a third of the time (32%) and most often required manual entry of data (54%). Studies of CDS for pediatrics and obstetric populations were absent.

The most common outcomes measured were process outcomes (79%). These included accurate interpretation by nurses of CDS information (25%), accuracy of CDS information (29%), situation awareness (32%), workload (21%), efficiency (21%) and errors (18%). Thirteen of the 22 studies measured >1 process outcomes. Four studies reported statistically significant improvements in processes; two for nurse accuracy and two for efficiency. Assessment of usability outcomes was also common (68%), including 4 types: global usability (21%), learnability (29%), subjective satisfaction (14%), and value/usefulness (32%). Statistically significant improvement was found in 14% of the studies measuring usability. The least common outcomes measured were patient outcomes (18%), of this small group of studies only two demonstrated statistically significant patient improvement.

**Discussion/Conclusions**

While CDS targeted to nursing decisions hold promise for improving patient care, this review found statistically significant improvement in outcomes in only 40% of the studies. Process measures improved for 4 studies, usability improved for 5 studies and patient outcomes improved in 2 studies. There are several reasons for the relatively small number. First, 15 included studies reported only descriptive or qualitative results. Second, for patient outcomes, there were 14 studies in which patient outcome measurement was not applicable (e.g. lab based studies). For the lab-based studies, it is not clear if the CDS are now used in practice and therefore suitable for patient outcome measurement. In summary, for nursing CDS to deliver on the promise of improved care, research must progress beyond the lab stage and evaluate measureable patient outcomes using inferential statistics and more sophisticated study designs.
Figure 1. Records Selection Process

1. Title review: Not in English (28), conference proceedings (644), not exclusively focused on R.N.s (750), not healthcare related (50), and not inpatient acute care (305)

2. Abstract review: Not research (86), a review (17), students as the only subjects (9), not only RNs (133), not automated as health information technology (98), not CDS in its current form (154), or CDS was not patient care related (18)

3. Article Review: Not research (4), students were the only subjects (1), not exclusively R.N. focused (4), not health information technology (2), not CDS (4), or the CDS was not related to caring for patients (2)

References

Automating Guidelines for Clinical Decision Support (CDS): A Categorization of Knowledge Engineering and Implementation Decisions

Mary K. Goldstein MD MS,1, 2 Samson W. Tu MS,2,1 Connie Oshiro PhD,1 Susana Martins MD MSc,1 Dan Wang PhD,1 Amy Furman PharmD,1 Michael Ashcraft MD,1 Jonathan Mendoza,1 Paul A. Heidenreich MD MS1,2, Kaeli Yuen,1 1VA Palo Alto Health Care System, Palo Alto, CA; 2Stanford University, Stanford, CA

Introduction: ATHENA-CDS [1] is a CDS system developed at VA Palo Alto Health Care System and Stanford University. ATHENA-CDS has encoded complex clinical practice guideline recommendations from five clinical domains: hypertension, chronic kidney disease, heart failure, hyperlipidemia, and diabetes. As part of a project to deliver CDS to health professionals for VA patients whose data do not meet one or more performance measures, the CDS is being integrated into an existing clinical Dashboard that flags patients not meeting performance measures. In this study, we seek to validate and to extend the 13 steps Shiffman [2] identified for translating the knowledge contained in guideline text into a computable format and to integrate the information into clinical workflow.

Methods: We identified 119 decisions made at project meetings over an 8-month period. One project member, an experienced knowledge engineer, evaluated each decision to categorize the decisions into Shiffman’s steps, and, when such categorization was not possible, proposed new categories. A senior clinician evaluated the initial categorization and, working with the knowledge engineer, developed a consensus categorization of the decisions.

Results:

Table 1. Frequency of decisions that correspond to Shiffman’s steps [2]

<table>
<thead>
<tr>
<th>Select Guidelines</th>
<th>Mark-up</th>
<th>Atomize</th>
<th>De-abstract</th>
<th>Disambiguate</th>
<th>Build Statements</th>
<th>Verify Completeness</th>
<th>Add Explanation</th>
<th>Identify Origin</th>
<th>Identify Insertion</th>
<th>Define Action Type</th>
<th>Define Associated Beneficial Services</th>
<th>Choose Interface Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Frequency of decisions that correspond to new categories

<table>
<thead>
<tr>
<th>Alignment with Existing Dashboard</th>
<th>Local Adaptation</th>
<th>Adherence</th>
<th>Terminology Mapping</th>
<th>Testing</th>
<th>Knowledge Representation</th>
<th>Multiple Guidelines</th>
<th>Project Management</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

(See Tables 3 and 4 for definitions of the categories.)

Discussion: Some of Shiffman’s steps, such as atomize, de-abstract, and disambiguate, are not reflected or are under-represented in ATHENA-CDS’s formally recorded decisions because project knowledge engineers perform them without requiring the whole team making decisions about them. A number of factors account for decisions that cannot be classified using Shiffman’s categories. First, ATHENA-CDS generates recommendations for patients who may have multiple co-morbidities, making interactions among guideline recommendations an important concern. Shiffman’s steps envision formalizing a set of recommendations within a system where the implementers can make their own decisions about such things as how to disambiguate concepts. In contrast, because ATHENA CDS is being integrated into an existing Dashboard, we have to take into account decisions that developers of the Dashboard have already made. We can also make use of data analytics (e.g., patients’ adherence to prescribed medications) that the Dashboard makes available. With the proliferation of health information technology (HIT) tools, this is likely to be a common scenario for CDS developers. ATHENA-CDS is implemented using a rich guideline model, not rules with decision variables that Shiffman assumes. Thus we have to make decisions about knowledge representation (e.g., how a domain ontology should be organized) that are not recognized in his methodology. Finally, ATHENA-CDS is a working system, not just an implementation specification that Shiffman’s methodology produces, making issues such as local adaptation, terminology mapping, testing, and project management important.

Conclusion: We validated and extended steps in Shiffman’s approach to making guideline recommendations computable. We identified additional knowledge engineering and implementation categories needed because of multiple comorbidities and guidelines, actual system implementation, and integration with existing HIT tools.
Table 3. Definitions of Shiffman's Steps (Adapted from [2])

<table>
<thead>
<tr>
<th>Select Guidelines</th>
<th>Choice of specific guidelines and choice of specific recommendations within the selected guidelines to be implemented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark-up</td>
<td>Identification and tagging of guideline knowledge components relevant to operationalization</td>
</tr>
<tr>
<td>Atomize</td>
<td>The process of extracting and refining single concepts from the recommendation’s natural language text</td>
</tr>
<tr>
<td>De-abstract</td>
<td>The process of adjusting the level of generality at which a decision variable or action is described to permit operationalization</td>
</tr>
<tr>
<td>Disambiguate</td>
<td>The process of establishing a single semantic interpretation for a recommendation statement</td>
</tr>
<tr>
<td>Build Executable</td>
<td>Arrangement of the atomized, de-abstracted, and disambiguated decision variables and actions into logical statements that can be translated readily into computable statements</td>
</tr>
<tr>
<td>Verify Completeness</td>
<td>The process to make sure that each recommendation provides guidance in all situations that a clinician is likely to face</td>
</tr>
<tr>
<td>Add Explanation</td>
<td>A facility to describe the reasoning behind recommendations</td>
</tr>
<tr>
<td>Identify Origin</td>
<td>Identifying a source or origin in the clinical environment for each decision variable</td>
</tr>
<tr>
<td>Identify Insertion</td>
<td>Identifying an insertion point in the care process for each recommended action</td>
</tr>
<tr>
<td>Define Action Type</td>
<td>Categorizing guideline-recommended activities according to predefined action types</td>
</tr>
<tr>
<td>Define Associated</td>
<td>Linking action types to associated beneficial services that offer design patterns for facilitating clinical care</td>
</tr>
<tr>
<td>Beneficial Services</td>
<td></td>
</tr>
<tr>
<td>Choose Interface</td>
<td>Selection and grouping user interface elements</td>
</tr>
<tr>
<td>Component</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Definitions of New Decision Categories

<table>
<thead>
<tr>
<th>Alignment with Dashboard</th>
<th>Choice of details of recommendations (e.g., HbA1C targets and drugs to recommend) based on the need to be consistent with the Dashboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Adaptation</td>
<td>Modification of recommendations based on local conditions (e.g., non-availability of data)</td>
</tr>
<tr>
<td>Adherence</td>
<td>Checking patients’ adherence to prescriptions based on medication possession ratios calculated by the Dashboard</td>
</tr>
<tr>
<td>Terminology Mapping</td>
<td>Mapping terminology used in data sources to terminology used in guidelines</td>
</tr>
<tr>
<td>Testing</td>
<td>Selection and running simulated and real patient data to verify correctness of guideline encoding and the completeness and clinical appropriateness of system-generated recommendations</td>
</tr>
<tr>
<td>Knowledge Representation</td>
<td>Decisions about versioning, the organization of, and conventions used in the encoded guideline knowledge bases. Also the representation choice of specific guideline knowledge.</td>
</tr>
<tr>
<td>Multiple Guidelines</td>
<td>Reconciliation of recommendations from different guideline sources (e.g., VA versus professional society guidelines) and guidelines for related clinical domains.</td>
</tr>
<tr>
<td>Project Management</td>
<td>Decisions about the scope of the project and the organization and timing of tasks</td>
</tr>
</tbody>
</table>

References:


Acknowledgement

Project supported by VA HSR&D grant IIR 11-071. Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs or other affiliated institutions
Crowdsourcing STAR annotations for large scale molecular classification of disease

Dexter Hadley¹, James Pan², Osama El-Sayed³, Jihad Al-Jabban⁴, Imad Al-Jabban⁴, Tej Azad², Shuaib Raza³, Mohamed Hadied⁵, Bin Chen¹, Hyojung Paik¹, Sanchita Bhattacharya¹, Marina Sirotai¹, Atul J. Butte¹

¹ Institute for Computational Health Sciences, University of California, San Francisco, CA 94158, ² Stanford University School of Medicine, Stanford, CA 94305, ³ University of Illinois College of Medicine, Chicago, IL 60612, ⁴ Harvard Medical School Department of Immunology, Harvard University, Boston, MA 02115, ⁵ Wayne State University School of Medicine, Detroit, MI 48201, ⁶ Yale School of Medicine, Yale University, New Haven, CT 06519

Abstract. There are over 126 billion functional genomics data points from more than 1.6 million digital samples in public functional genomics experiments amassed over the last 15 years. This highest-quality, NIH-funded functional genomics Big Data can be translated into large-scale drug and biomarker discovery as our lab and others have repeatedly shown. However, individual samples remain poorly described by unstructured free text that represents complex bio-ontologies that are shared by samples across independent experiments. Thus far, strictly computational approaches to annotate GEO en-masse such as text mining have largely failed to recapitulate biological annotations with any measured reliability. Therefore, to address this growing Big Data problem and to scale the precise annotation and useful interpretation of open digital samples, we previously built the Search Tag Analyze Resource (STAR; stargeo.org) to allow anyone to describe GEO sample phenotypes uniformly across different studies and to characterize gene expression signatures for disease. Here, we design and implement a validation system based on STAR annotations to reward concordance among expert inter-raters and converge on precision annotations at scale. Furthermore, we use meta-analytics across many thousands of samples with shared phenotypes to estimate a molecular nosology, or classification of disease, for over 20 different autoimmune, infectious, or neoplastic clinical conditions. Our innovative approach uses emerging Internet technologies such as crowdsourcing and social networks to translate open Big Data in biomedicine into structured knowledge that can be used at scale for massive drug and biomarker discovery.

Introduction. The NCBI’s Gene Expression Omnibus (GEO) is perhaps the largest of a number of public data repositories¹⁻³, and it currently shares the gene expression measurements of over 1.6 million samples drawn from functional genomics over the last 15 years.⁴⁻⁶. The term ‘crowdsourcing’ was first described in 2006,⁷ and it describes the process of breaking a complex problem into smaller ‘micro-tasks’ that engage a crowd of participants to jointly undertake. Crowdsourcing is rapidly being applied to bioinformatics problems that require human involvement⁸⁻¹⁰, and colleagues have had recent success in using crowds to annotate PubMed abstracts for mentions of disease and genes with non-specialized curatorsⁱ¹, while others have had mixed results making biological annotations on unstructured text with a naive or uninformed crowd.¹². We previously developed the Search Tag Analyze Resource (STAR) web application¹³ (http://stargeo.org) to leverage crowdsourcing among biomedical curators to annotating GEO.¹⁴ The STAR annotation process included a web interface for curators to design regular expressions (RegExs), a standardized syntax in computer science,¹⁵ to efficiently match and extract text, as shown in in Figure 1.

Figure 1: STAR annotation process. The flow chart shows how unstructured free text attributes from GEO are annotated by specialized curators across Tags. The crowd repeatedly makes multiple passes at the annotation until at least two independent experts agree. The figure also shows a screenshot of the STAR annotation process for Tagging the experimental study (GSE12351) entitled “Melanoma: comparison between common nevi, radial-vertical growth phase melanoma, metastases and dysplastic nevi”. The melanoma Tag is assigned a RegEx that maps unstructured ‘melanoma’ free text attributes to GEO sample descriptors. STAR automatically highlights matching samples based with by real-time RegEx search. Therefore, by entering only the unstructured text attributes that discriminates ‘melanoma from ‘common acquired melanocytic nevus’ as shown in this example, a curator can quickly annotate disease, outcome, phenotype or any other experimental qualifier. This process is repeated across different studies.
Methods. We reimbursed $10,000 to curators at a rate of 10c for every concordant sample annotated with 5c apiece going to each curator. We integrated reimbursement by electronic gift cards directly into the STAR web interface using the Tango Card (http://tangocard.com) API for access to over 70 retailers including Amazon, Wal-Mart, and Starbucks. We used meta-analysis performed by STAR to estimate a nosology for over 20 diseases whose annotations were crowdsourced by biomedical experts. For each gene across every disease phenotype, we calculated the significance of effect of that gene in disease as the inversely scaled estimate of statistical significance of the mean expression difference of cases and controls assuming a random effects model. We then used agglomerative clustering significance plot a dendrogram as a molecular classification for disease or nosology.

Results. We used Twitter to strategically recruit STAR curators that would be interested in learning about disease and defining genomic disease signatures by querying key words like “biomedicine”, “translational medicine”, “research”, etc. In 8 months, we have acquired approximately >485,000 annotations, and we have demonstrated high (>90%) precision among curators that cross annotated each others efforts (Figure 2). Current projections suggest that within three years we can achieve complete coverage of GEO data with at least one STAR annotation (Figure 3). Therefore, STAR crowdsourcing may be the only way to completely annotate GEO. We clustered all the diseases analyzed by the significance of their estimated meta effects over 11,267 genes that were measured across all samples to estimate a molecular nosology of 20 different disorders (Figure 4).

Discussion. STAR represents an annotation layer built on GEO data, and it provides convenient tools for biomedical experts to derive robust gene signatures from open big data. Here, we implement a validation system that converges on precision of annotations at scale. We do so by rewarding concordance among expert raters. Furthermore, we use machine learning of meta-effects across many thousands of samples with shared phenotypes to estimate a molecular nosology, or functional classification of disease. Our innovative approach uses emerging Internet technologies such as crowd-sourcing and social networks to translate open Big Data in biomedical into structured knowledge that can be used at scale for subsequent large-scale drug and biomarker discovery.

The nosology we estimate very cleanly segregates cancer (PAC=pulmonary adenocarcinoma, HCC= hepatocellular carcinoma, HNSCC=Head and neck squamous cell carcinoma, Lymphoma, Medulloblastoma) from other non-neoplastic disorders. Moreover, this functional representation of disease may suggest clinical utility to repurpose drugs indicated for one disease to another functionally related disease. A good example in our current tree is the proximity of HCC and PAC suggests to consider shared treatments as a potential therapeutic option for either.

Conclusion. Our results prove the translational utility of STAR as an online platform and the feasibility of using crowdsourcing as a paradigm to completely annotate GEO within three years. As we continue to extend our STAR platform, we demonstrate how Internet innovations such as social networks can be used to help characterize disease using open functional genomics Big Data. By estimating a robust and precise molecular nosology of disease, our work allows for massive and systematic drug and biomarker discovery in the near future. Therefore, we are defining the utility of crowdsourcing to translate open biomedical Big Data into useful medical innovation.

Acknowledgements. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01 GM079719, and the National Institute of Allergy and Infectious Diseases (Bioinformatics Support Contract HHSN272201200028C). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Figure 2: STAR annotations. The figure shows the cumulative distribution of 485,054 total annotations over nine months with >90% concordant (green) and <10% discordant (red) annotations that were performed twice, relative to the sample annotations that were only performed once (gray). Blue dashed box labels the effect of $10,000 reimbursement to yield over 360,000 biological annotations in only 6 weeks.

Figure 3: STAR sample annotation. The graph shows the cumulative accumulation of (blue) GEO samples vs coverage with (red) STAR annotations. Linear models (dashed lines) project complete STAR coverage of GEO within three years at current rates.

Figure 4: Tree of disease. The figure shows the clustering dendrogram across 12767 genes across 20 different diseases. For each disease, the number of studies (GSEs), genomic platforms (GPLs), and samples (GSMs) is listed.
Making nursing visible in health information systems

Nicholas R. Hardiker, RN PhD FACMI¹, Nicola Fortune, BSc², Kay Jansen, RN DNP³, Claudia C. Bartz, PhD, RN, FAAN³
¹University of Salford, UK; ²University of Sydney, Australia; ³University of Wisconsin-Milwaukee

Introduction

Nurses are the largest group of health workers in most health care systems. While data-based information can be used to characterize nursing care activities and costs, and to demonstrate the value of nursing, the actual use of information for these purposes has been patchy, making local, national or international comparisons difficult, if not impossible. The World Health Organization’s (WHO) International Classification of Health Interventions (ICHI) is being developed as an international standard for collecting, reporting and analyzing data on health interventions. ICHI spans all sectors of the health system and a wide range of potential applications are envisaged, including use of the classification as a building block for international casemix development. Inclusion of nursing interventions in ICHI is essential to ensure that nursing is represented in the international health information infrastructure of the future. The International Council of Nurses (ICN) is contributing to the development of ICHI, using the International Classification for Nursing Practice (ICNP), a Related Classification within the WHO Family of International Classifications, as a resource for nursing content. The work described in this paper comprises a comparison of content and a gap analysis, using ICNP as a source to identify potential additional content for ICHI.

Methods

There were two phases to the work, centering around two distinct sub-sets of ICNP: a) a set of interventions that were subsumed within the OWL representation of ICNP by the intervention types informing, determining, managing and performing (as more indicative of direct care actions) and b) a much broader set of the remainder of ICNP interventions. In both phases approximate initial candidate mappings between ICNP and ICHI were identified by a member of the ICHI team. Where no mapping could be found, the construction of new interventions by combining relevant ICHI Target and Action categories was suggested (and new categories proposed as necessary). Members of the ICNP team provided advice as to whether ICNP interventions were conceptually equivalent or represented a subclass (more detailed concept) or superclass (broader concept) in relation to the suggested ICHI intervention, using the standard practice of independent review with follow-up consensus work.

Results

In the first phase, potential mappings were identified for 480 ICNP interventions. Reviewers agreed on 324 mappings from ICNP to ICHI (67.5%). Proposals were made to the ICHI development team for the inclusion of new ICHI content, which resulted in sixty-seven new nursing-relevant interventions and seven new Target categories. In the second phase, potential mappings were identified from the remaining 337 ICNP interventions to ICHI. There were 183 additional agreed mappings from ICNP to ICHI (54.3%). This second phase has resulted in further proposals to the ICHI development team, which will inform the inclusion of additional ICHI content.

Discussion

In total there were 817 suggested mappings from ICNP to ICHI, of which 507 were agreed (62.1%). There are a number of reasons behind the lack of mappings for the remaining ICNP interventions, including a) differences in scope between ICHI and ICNP - the focus for ICHI is primarily on direct care activities while ICNP takes a broader view of nursing interventions (a possible explanation for the lower proportion of mappings for the second phase), and b) the absence of ICHI Target categories able to adequately describe some ICNP interventions. This work, a comparison of content and gap analysis, has served as a useful starting point for a larger program of research around nursing and casemix involving an exploration of a) the ability of ICHI to capture existing data-based ICNP-encoded data, in order to concentrate attention on those interventions that nurses actually perform in practice, rather than on a potentially much larger group of interventions that nurses might perform, and b) the potential contribution of ICHI to existing tried-and-tested systems for nursing resource management.

WHO estimates that there are 29 million nurses and midwives across the world; however their contribution remains largely hidden within today’s health information systems. The foundational work described in this paper, and the ongoing research, seeks to help expose the true value (and true cost) of nursing.
A survey of social media for understanding patient-reported medication outcomes

Kimberly Harris, MS\textsuperscript{1}, Boshu Ru, MS\textsuperscript{1}, Lixia Yao, PhD\textsuperscript{1}

\textsuperscript{1}Department of Software and Information Systems, The University of North Carolina at Charlotte, Charlotte NC 28223

**Introduction:** Pre-marketing clinical trials provide limited perspectives about new drugs’ safety and risks, due to marginal patient enrollment and short-term experiments. It is essential to continuously monitor medication outcomes among large and diverse patient populations, in order to thoroughly understand drugs’ long-term effectiveness, side effects, and life impacts [1,2]. Failure to do so may cause serious morbidity or even mortality (e.g., the withdrawal of Vioxx [3]). In response, the United States Food and Drug Administration (FDA) maintains the Adverse Event Reporting System (FAERS) to collect adverse drug events from clinicians, pharmaceutical companies, and patients. However, this system is often criticized for its low reporting rate, undesirable data quality, and potential conflict of interest [4]. In contrast, there is an emerging trend by consumers to discuss their diagnoses and treatments with other people that have similar symptoms and ailments on popular social media platforms, such as PatientsLikeMe and Twitter. In this project, we plan to investigate whether social media can be a complementary data source for medication outcomes research.

**Methods:** We selected four chronic diseases, namely asthma, rheumatoid arthritis, type 2 diabetes, and cystic fibrosis. For each of these diseases, we picked two to three commonly prescribed drugs. Thus, for the total of 11 disease-drug pairs, we investigated three particular questions: (1) do consumers report their medication outcomes and experiences on social media? (2) what social media platform is the best to study consumer-reported medication outcome, according to data volume and quality? (3) what challenges are there to fully analyze the textual data shared by consumers on social media platforms? We selected two professional health social media sites and two general-purpose social media sites as our data sources. The selection of each social media site took into account both its prevalence in the population and data accessibility. The user comments for the 11 disease-drug pairs were downloaded from WebMD, PatientsLikeMe, YouTube, and Twitter. Disease and drug synonyms were used in the search to ensure the completeness of the results. Afterwards, all users’ posts were put into one of four categories: effectiveness, side effect, adherence, and cost at the sentence level, using terminologies, such as Consumer Health Vocabulary. In the end, we did the sentiment analysis using the Stanford sentiment analysis tool [5].

**Results:** Table 1 summarizes the number of user posts from WebMD, PatientsLikeMe, YouTube, and Twitter for each disease-drug pair. The results demonstrate that consumers did post a lot of information on social media relevant to medication outcomes for all four diseases. Though there were fair amounts of relevant text content posted on each social media platform, WebMD contained the richest information regarding medication outcomes among the four. Both WebMD and PatientsLikeMe provided rating schemas to reflect users’ opinions of medication outcomes in a quantitative manner. The Pearson’s correlation coefficient equaled to 0.728 and demonstrated consistency in user reports of drug effectiveness for these two sites. On PatientsLikeMe, however, only 3% of the users described their medication experiences in free text. Majority of the posts on YouTube and Twitter contained irrelevant content, such as drug related news, commercial advertisement, and clinical study or publication announcements, with few relating to medication outcomes and experiences. Figure 1 illustrates the sentiment analysis results associated with the users’ comments on different medication outcome topics. Negative comments and side effects dominated on all four social media platforms.

**Discussion:** While reviewing the comments retrieved, we found a few users discussed how the drugs they were taking unexpectedly helped their secondary conditions, which exemplified potential opportunities for drug repositioning [6]. Users on WebMD and PatientsLikeMe appear to be more conscious and knowledgeable about their drug therapies. Their feedback is usually more informative than those on YouTube or Twitter, which often lack context. Also, majority of the posts consists of adult-results, and not those reflective of teenagers or children. It will be interesting to see how and where the younger generation posts their observations in future studies. Furthermore, consumers wrote with language that is challenging for computers to process, such as sarcasm, grammatical errors, typographical facial expressions, comparative opinions about different drugs, and various phrases for analogous attitudes. Despite those challenges, we believe social media can be a complementary new data source because of its high prevalence in the population and its independence from other data sources, which enables us to use it to validate the patterns identified in the traditional data sources. Our future works will focus on implementing natural language processing and text mining methods suitable for using social media to systematically mine medication outcome contents on a large scale.
References

Table 1: Summary of user comments among the four social media sites

<table>
<thead>
<tr>
<th>Social Media and Metrics</th>
<th>WebMD</th>
<th>PatientsLikeMe</th>
<th>YouTube</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases and Drugs</td>
<td>Effectiveness</td>
<td>Ease of Use</td>
<td>Satisfaction</td>
<td>No. of Reviews</td>
</tr>
<tr>
<td>Asthma</td>
<td>Albuterol 3.87</td>
<td>4.39</td>
<td>3.55</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>Ipratropium 4.17</td>
<td>4.17</td>
<td>3.75</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Prednisone 4.00</td>
<td>3.92</td>
<td>3.32</td>
<td>367</td>
</tr>
<tr>
<td>Cystic Fibrosis</td>
<td>Azithromycin -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Ivacaftor -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>Meloxicam 3.39</td>
<td>4.18</td>
<td>3.11</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td>Prednisone 4.11</td>
<td>4.31</td>
<td>3.61</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>Sulfasalazine 3.19</td>
<td>3.28</td>
<td>3.17</td>
<td>65</td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>Bromocriptine 2.23</td>
<td>3.08</td>
<td>2.15</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Insulin 3.50</td>
<td>4.22</td>
<td>3.35</td>
<td>265</td>
</tr>
<tr>
<td></td>
<td>Metformin 3.29</td>
<td>3.91</td>
<td>2.93</td>
<td>1302</td>
</tr>
</tbody>
</table>

See Figure 1: Sentiment analysis on user comments across the four social media sites.
Impact of Electronic Health Records on Quality of Care: Evidence on Inpatient Mortality, Readmissions, and Complications

Tina Hernandez-Boussard, Doug Morrison, Swati Yanamadala, Catherine Curtin, Kathryn McDonald

Research Objective: It has been suggested that health information technology, particularly electronic health records (EHR), will improve quality and efficiency of healthcare organizations and many hospitals have implemented these systems. However, little evidence exists on the association between EHR adoption and improved quality of care across a broad range of medical and surgical conditions. We hypothesized that hospitals with fully-implemented EHR-systems [Full-EHR] would have lower levels of inpatient mortality, 30-day all cause readmissions, and complications (measured by Patient Safety Indicators [PSIs]), compared to hospitals with no or partially implemented EHR-systems [No-EHR and Partial-EHR, respectively].

Study Design: We used data from the Healthcare Cost and Utilization Project (HCUP) and the American Hospital Association annual surveys (AHA) from California, Florida, and New York, 2008-2011. Cross-sectional random-effect logistic regression models were used to determine associations between EHR implementation level and quality of care. Relative-risk difference in differences analyses (DiD) were used to determine the effect of implementing an EHR-system on quality of care.

Population Studied: We included surgical patients undergoing coronary artery bypass graft, pulmonary lobectomy, open abdominal aortic aneurysm repair, endovascular abdominal aortic aneurysm repair, or colectomy (n=202,936) and medical patients receiving care for acute myocardial infarction, coronary heart failure, and pneumonia (n=439,410).

Principal Findings: Medical and surgical patients sought care at hospitals reporting No-EHR (10.1%), Partial-EHR (56.5%) and Full-EHR (33.4%). Patients at hospitals with a Full-EHR had the lowest rates of inpatient mortality, readmissions, and PSIs followed by patients at hospitals with a Partial-EHR and then patients at hospitals with No-EHR. (p<.05); surgical patients demonstrated between a 9-15% decrease and medical patients had between a 5-12% decrease in rates. Adjusting for patient and hospital characteristics, both medical and surgical patients at hospitals with Full-EHR had between 7-24% lower relative risk for inpatient mortality, readmission, and PSIs compared to patients at hospitals with No-EHR and Partial-EHR. However, the DiD found EHR implementation had a minimal effect both in medical and surgical patients on inpatient mortality readmission, and complications.

Conclusions: Patients seeking care at hospitals with either a fully implemented or partially implemented EHR-system had lower rates of inpatient mortality, readmission, and PSIs compared to patients at hospitals with No-EHR. These benefits were significant for both surgical and medical patients and varied by
specific medical condition and procedure. However, the DiD analysis suggests that factors other than EHR implementation account these differences in quality.

**Implications for Policy and Practice:** As federal incentives encourage EHR adoption and hospitals strive for meaningful use, it will be important to understand the benefits received from EHRs. Under the healthcare reform, all parties focus on the quality of care received and EHRs might play a smaller role than expected. Our results suggest that EHR-associated improved quality, safety, and efficacy might vary by type of care and meaningful use targets should reflect such variations.
Discharge Instructions: What Do Patients Remember?
Brent Hill, RN MS PhD, Qing Zeng-Treitler, PhD
University of Utah, 421 Wakara Way, Ste 140, Salt Lake City, Utah 84108

Introduction When a patient leaves the hospital, he is provided with instructions for his post-hospitalization care, follow-up appointments and tests, and other patient education information. However, patients have a difficult time understanding their discharge instructions.[1-3] The objective of this study is to understand the types of information patients remember and do not remember in their discharge instructions. This information will be used to inform informatics interventions to improve patient comprehension, recall and adherence to discharge instructions. This study builds on prior research in which illustrated versus non-illustrated discharge instructions were tested with 144 patients to evaluate their recall of and satisfaction with their discharge instructions.[4, 5] The first 50 discharge instructions were analyzed for this pilot study.

Methods Patients on a Cardiovascular Medical Unit at the University of Utah Hospital were asked to remember the content of their disease or procedure specific discharge instructions after their final discharge teaching was completed by their nurse and they were discharged to go home. Study participants were asked to review their instructions for up to 15 minutes then prompted by the study nurse to recall their instructions as the nurse asked them to state all of the information they could remember as the nurse read the title of each discharge instruction section such as diagnosis, signs/symptoms to be vigilant for, wound care, activity restrictions until each section was reviewed. Finally, the patient was asked for any additional information they recalled. Three levels of analysis were conducted for this study. The first level of analysis consisted of reviewing discharge instructions by section. The second level of analysis consisted of reviewing discharge instructions in detail by individual instruction. The third level of analysis evaluated the information based on the immediacy of the need for the instruction. Patient’s discharge instructions were based on their diagnosis or procedure so instructions varied. Content analysis was completed by quantification of instructions by remembered or did not remember and a ratio was calculated based on the percentage of participants who received that instruction who did or did not remember the instruction.

Results The results for the first level of analysis revealed the percentages patients did/did not remember in the following sections of discharge instructions: Diagnosis (78/22), General Care (17/83), Wound Care (26/74), Activity Level (22/78), Restrictions (39/61), Diet (24/76), Other (4/96) (Table 1). The second level of analysis presents selected discharge instructions and the percent of patients that did and did not remember the instruction (Table 2). The third level of analysis examined the immediacy of instructions versus general health education. A cardiologist independently evaluated the discharge instructions for immediacy of information by highlighting information that he deemed necessary for the patient’s first week post-hospitalization. Analysis of the instructions revealed that 27% of the discharge instructions were necessary for 1 week post-hospitalization and 73% of the discharge instructions consisted of general health education.

Discussion It is necessary to understand how patients retain information from discharge instructions in order to prioritize the content and delivery of discharge instructions. Suggested improvements include: simplify text, stratify information based on need, and use different delivery methods such as traditional paper, electronic via email or PHR, and utilize info graphics or pictographs. This information will be used to inform informatics interventions that will tailor and deliver information for the patient that can support the patient during this transition of care.
Table 1. Percent of Discharge Instructions by Section.

<table>
<thead>
<tr>
<th></th>
<th>Primary Diagnosis</th>
<th>General Care</th>
<th>Wound Care</th>
<th>Activity</th>
<th>Restrictions</th>
<th>Diet</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remembered</td>
<td>78%</td>
<td>17%</td>
<td>26%</td>
<td>22%</td>
<td>39%</td>
<td>24%</td>
<td>4%</td>
</tr>
<tr>
<td>Did Not Remember</td>
<td>22%</td>
<td>83%</td>
<td>74%</td>
<td>78%</td>
<td>61%</td>
<td>76%</td>
<td>96%</td>
</tr>
</tbody>
</table>

Table 2. Percent of Discharge Instructions by Instruction.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Remembered</th>
<th>Did Not Remember</th>
</tr>
</thead>
<tbody>
<tr>
<td>Having signs of a stroke including weakness or numbness in your face or</td>
<td>9%</td>
<td>91%</td>
</tr>
<tr>
<td>limbs-especially on one side of the body- confusion, loss of balance,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trouble speaking or seeing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Look at your wound every day for signs of infection.</td>
<td>26%</td>
<td>74%</td>
</tr>
<tr>
<td>If you start to bleed from an incision, apply pressure for 15 minutes.</td>
<td>34%</td>
<td>66%</td>
</tr>
<tr>
<td>Do not bear down or push hard to have a bowel movement or strain while</td>
<td>6%</td>
<td>94%</td>
</tr>
<tr>
<td>working.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What Are Frequent Data Requests from Researchers? A Conceptual Model of Researcher’s EHR Data Needs for Comparative Effectiveness Research

Gregory W. Hruby¹, MA, Praveen Chandar¹, PhD, Julia Hoxha¹, PhD, Eneida A. Mendonça²,³, MD, PhD, David A Hanauer⁴,⁵, MD, MS, Chunhua Weng, PhD¹
¹Dept. of Biomedical Informatics, Columbia University, New York, NY; ²Dept. of Pediatrics; ³Dept. of Biostatistics and Medical Informatics University of Wisconsin, Madison, WI; ⁴Dept. of Pediatrics, ⁵School of Information, University of Michigan, Ann Arbor, MI

Introduction

Data request forms are the key communication media linking medical researchers and informaticians¹. A recently published Carpenter framework organizes data needs for cancer comparative effectiveness research². It complements the PICO framework that defines the Patient, Intervention, Comparison, and Outcome constructs for specifying clinical information needs³-⁵. Our goals are to validate, enrich and generalize the Carpenter framework utilizing three data sources representing researcher information or data needs. We posit the Carpenter model may serve a standard template for data needs specification as it contains semantic structural similarities to the PICO framework. We redefined the model as a representation of abstract medical concepts used in research. We extended the Carpenter framework by expanding the model’s breadth to multiple medical research domains.

Methods

We utilized three data sources for this study: Clinical trial inclusion/exclusion criteria, EHR SQL project queries, and EHR data request logs. Each source represents a unique aspect of data needs specification across different research needs and different medical disease domains. For example, clinical trial inclusion/exclusion criteria provide ideal data needs specification for the definition of cohort identification across a diverse set of diseases while the EHR SQL project queries provide a comprehensive list of explicit granular concepts organized within the context of real world data needs used for retrospective CER of benign and malignant urologic diseases.

We parsed each dataset to the sentence/variable level. From the clinical trial inclusion/exclusion criteria and the EHR data request logs, we randomly selected 1,000 and 897 sentences, respectively. From the SQL project files, we extracted 1,445 distinct variables. Each dataset was iteratively annotated with the Carpenter framework by one coder (GH). Nodes were removed or added to the Carpenter framework when the concept was not used, or the concept did not previously exist, respectively. The proposed model was created with the enriched concept list. Domain experts evaluated if the functionality of the enriched model meets the intended purpose: a guideline to aid the specification/elicitation of EHR data needs used in research.

Results

Figure 1 represents the enriched framework. Six original child nodes were pruned from the Carpenter framework. Pruned nodes, e.g. Local disease burden, health norms, care process guidelines, and care systems and coordination, are not well represented in the EHR and as such, many medical researchers would forgo any mentioning of these concepts. Fifteen additional leaf nodes were added to the representation and are shown by an underlining of the concept within the figure. The Organizational/Provider Characteristics parent node (the boxes) was moved to the second column in the proposed model, as the data suggested it was more related to the parent nodes, Detection/Diagnostics and Intervention, rather than the Patient node. For the attribute of completeness, the model correctly identified 88% (63/72) of the medical concepts generated by our expert evaluators. For the attribute of generalizability, the model accurately identified medical concepts from diverse medical domains with a median accuracy rate of 91% (60-100%). For the attribute of reusability, the enriched model reused 68% (53/78) of the concepts from the original Carpenter model.

Discussion

This study contributes an enriched conceptual model representing the medical concepts used in clinical research. The major contribution of the enriched model is not the granular nodes, but it’s representation of the researcher’s conceptual organization of medical concepts used for research. This may serve as a bridge between the medical researcher and the informatician by aligning the informatician’s mental model to researcher’s. Both stakeholders may use this model to specify and elicit key medical concepts needed for the research project.
Figure 1. The enriched model. The blue directed edges represent a temporal process as the patient moves through the care continuum. The cyclical nature of this graph implies the patient can re-enter the care cycle. The undirected edges indicate an association between the modules. The boxes represent parent nodes and the concepts within the boxes are the child nodes related to the parent nodes.

Acknowledgements
This study was sponsored by the U.S. National Library of Medicine grant R01LM009886 (PI: Weng) and U.S. National Center for Advancing Translational Science grant UL1 TR000040 (PI: Ginsberg).

References
Neighborhood Internet Access and Patient Portal Use in Patients with Chronic Conditions

Jie Huang PhD, Richard Brand PhD, Margaret Warton MPH, Bruce Fireman MA, John Hsu MD MBA MSCE, Ilana Graetz, PhD, Mary Reed DrPH

(1)Kaiser Permanente Division of Research, Oakland, CA, (2)Department of Epidemiology and Biostatistics, University of California at San Francisco, San Francisco, CA, (3)Mongan Institute for Health Policy, Massachusetts General Hospital, Harvard Medical School, Boston MA, (4)University of Tennessee Health Science Center, Memphis TN.

Introduction: Patient portals can improve patient engagement and health among patients with chronic diseases, but uptake and use of portals differs across racial/ethnic and socioeconomic groups. This study aims to examine the association between level of internet access, a possible underlying cause of the observed differences, and patient portal use among patients with chronic conditions in a large integrated health system.

Methods: In an integrated health delivery system with a web-based patient portal since 2005, patients can access personal medical information and interactive tools to manage health care at no charge, including viewing lab results and visit summaries, secure email messaging with health care providers, medication refill orders, and wellness programs. We studied use of the portal in 2013 by adult patients (aged 18 or older) in the chronic disease registry for risk of cardiovascular events (including diabetes, coronary artery disease, stroke, abdominal aortic aneurysm, and peripheral vascular disease) who were also continuously enrolled in the health system in the previous year. Level of internet access was measured by federal data for percent of households with a residential fixed high-speed internet connection in the census tract of the patient’s residence, categorized as low (<60% of households have internet connection), medium (60-80%) and high (>80%). We used logistic regression models to examine the association between level of internet access and portal use in 2013 (any portal use, lab result review, and secure messaging), adjusting for age, gender, neighborhood socioeconomic status, race/ethnicity, number of comorbid conditions in 2012 (as a measure of clinical need), history of medication adherence (as a measure of patient engagement), and medical center, and then used model results to calculate adjusted rates.

Results: Among 273,815 patients, 56.9% were 65 or older, 45.8% had non-white race/ethnicity, 21.0% lived in a low socio-economic-status neighborhood, and 49.2% were not adherent to previously prescribed chronic medications (proportion of days covered by medication <80%). Overall, 10.5% of all patients lived in a neighborhood with low internet access, 40.8% with medium, and 48.7% with high levels of internet access. Among all patients, 64.7% had used the patient portal for any tool (56.3% used the tool for lab results review and 58.6% used the tool for secure messaging) in 2013. Descriptively, use of the portal varied among patients with different level of internet access (low: 52.0%; medium: 61.4%; high: 70.2%). After adjusting for patient socio-demographic and clinical characteristics (Table), patients with higher level of internet access were more likely to use the portal overall (among low access: 58.1% used the portal, 95%CI: 57.4%-58.8%; among medium access: 62.1%, 95% CI: 61.8%-62.5%; among high access: 68.4%, 95%CI: 68.0%-68.7%). Those with higher internet access were also likely to use the tool for lab results review (low: 49.6%, 95% CI: 48.9%-50.3%; medium: 53.5%, 95% CI:53.2%-53.9%; high: 60.0%, 95% CI: 59.6%-60.4%) and more likely to use the tool for secure messaging (low: 51.7%, 95% CI: 50.9%-52.4%; medium: 56.1%, 95% CI: 55.7%-56.5%; high: 62.3%, 95% CI:61.9%-62.7%) . These patterns were consistent in analyses restricted to patients who had registered to use the portal.

Discussion: Patients with higher levels of neighborhood-level internet access were more likely to use patient portals, including specific lab result review and secure messaging tools, even after adjusting for patient socio-demographic characteristics, clinical need and patient engagement. The findings suggest that lack of home internet access may be a barrier to patient portal use. Future research should compare neighborhood-level measures with individually-reported access measures, including through mobile devices.
Table. Adjusted Association between Neighborhood Internet Access and Patient Characteristics and Portal Use

<table>
<thead>
<tr>
<th>Neighborhood Internet Access</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>Adjusted Percent</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>1.00</td>
<td>Ref.</td>
<td>58.1%</td>
<td>57.4%</td>
</tr>
<tr>
<td>Medium</td>
<td>1.20</td>
<td>1.16</td>
<td>1.24</td>
<td>62.1%</td>
</tr>
<tr>
<td>High</td>
<td>1.63</td>
<td>1.57</td>
<td>1.69</td>
<td>68.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient Age</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>Adjusted Percent</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-34</td>
<td>1.00</td>
<td>Ref.</td>
<td>84.1%</td>
<td>83.2%</td>
</tr>
<tr>
<td>35-44</td>
<td>0.79</td>
<td>0.73</td>
<td>0.86</td>
<td>81.0%</td>
</tr>
<tr>
<td>45-54</td>
<td>0.51</td>
<td>0.47</td>
<td>0.55</td>
<td>74.1%</td>
</tr>
<tr>
<td>55-64</td>
<td>0.46</td>
<td>0.43</td>
<td>0.50</td>
<td>72.3%</td>
</tr>
<tr>
<td>65-74</td>
<td>0.30</td>
<td>0.28</td>
<td>0.32</td>
<td>63.8%</td>
</tr>
<tr>
<td>75+</td>
<td>0.16</td>
<td>0.15</td>
<td>0.17</td>
<td>49.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>Adjusted Percent</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1.00</td>
<td>Ref.</td>
<td>65.7%</td>
<td>65.4%</td>
</tr>
<tr>
<td>Female</td>
<td>0.90</td>
<td>0.89</td>
<td>0.92</td>
<td>63.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race/ethnicity</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>Adjusted Percent</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>1.00</td>
<td>Ref.</td>
<td>73.0%</td>
<td>72.7%</td>
</tr>
<tr>
<td>Black</td>
<td>0.37</td>
<td>0.35</td>
<td>0.38</td>
<td>51.6%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.33</td>
<td>0.32</td>
<td>0.34</td>
<td>49.1%</td>
</tr>
<tr>
<td>Asian</td>
<td>0.59</td>
<td>0.57</td>
<td>0.61</td>
<td>62.2%</td>
</tr>
<tr>
<td>Other</td>
<td>0.42</td>
<td>0.40</td>
<td>0.45</td>
<td>54.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neighborhood SES</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>Adjusted Percent</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>1.00</td>
<td>Ref.</td>
<td>66.6%</td>
<td>66.3%</td>
</tr>
<tr>
<td>Low</td>
<td>0.68</td>
<td>0.66</td>
<td>0.70</td>
<td>58.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient Engagement</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>Adjusted Percent</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>1.00</td>
<td>Ref.</td>
<td>60.9%</td>
<td>60.2%</td>
</tr>
<tr>
<td>Medium</td>
<td>1.10</td>
<td>1.06</td>
<td>1.14</td>
<td>62.9%</td>
</tr>
<tr>
<td>High</td>
<td>1.29</td>
<td>1.25</td>
<td>1.34</td>
<td>66.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinical Needs</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>Adjusted Percent</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Low</td>
<td>1.00</td>
<td>Ref.</td>
<td>59.3%</td>
<td>58.7%</td>
</tr>
<tr>
<td>Low</td>
<td>1.19</td>
<td>1.16</td>
<td>1.23</td>
<td>63.0%</td>
</tr>
<tr>
<td>Medium</td>
<td>1.44</td>
<td>1.39</td>
<td>1.48</td>
<td>66.8%</td>
</tr>
<tr>
<td>High</td>
<td>1.48</td>
<td>1.43</td>
<td>1.53</td>
<td>67.4%</td>
</tr>
</tbody>
</table>
Usability Testing of an Ambulatory Navigator

Gretchen Hultman MPH\textsuperscript{1}, Elliot Arsoniadis MD\textsuperscript{1,2}, Jenna Marquard PhD\textsuperscript{3}, Rubina Rizvi MBBS, MS\textsuperscript{1}, Saif Khairat PhD\textsuperscript{4}, Keri Fickau\textsuperscript{5}, Genevieve B. Melton MD, PhD\textsuperscript{1,2}

\textsuperscript{1}Institute for Health Informatics, University of Minnesota, Minneapolis, MN
\textsuperscript{2}Department of Surgery, University of Minnesota, Minneapolis, MN
\textsuperscript{3}College of Engineering, University of Massachusetts, Amherst, MA
\textsuperscript{4}Carolina Health Informatics Program, University of North Carolina, Chapel Hill, NC
\textsuperscript{5}Fairview Health Services, Minneapolis, MN

Introduction

Although EHRs are widely used in clinical practice, poor user interface (UI) design continues to be a significant barrier to effective use of these systems for clinical care. These design problems can be tied to poor system usability. One approach to make EHRs more usable is to follow user-centered design (UCD) practices, which often includes formal usability testing. In response to clinician problems and challenges with a navigator in a new EHR, individuals at a large health system redesigned an ambulatory navigator within the EHR. Usability evaluations of two versions of an ambulatory navigator were conducted to understand if the new navigator improved user performance and satisfaction.

Methods

The original navigator (Figure 1) consisted of a column of links divided into domains. The redesigned navigator (Figure 2) was created in an iterative process in which clinicians from different specialties were directly involved in providing design guidance and testing of the navigator and had opportunity to provide guidance and feedback. In the final design of the new navigator, there are two buttons called “Intake” and “Charting”, and clicking either of these tabs show options pertaining to nursing/clinic staff and provider roles. The evaluation of the navigators utilized an EHR training environment replicating the production EHR with synthetic standard patients. Five patients of similar complexity were selected. Users completed a set of tasks for each patient, designed to align with Meaningful Use Stage 2 requirements. Tasks were first pilot tested with two residents and found to be similar in complexity, taking a similar amount of time to complete and being subjectively reported as similar difficulty. A convenience sample of residents (n=8) was recruited for this study. Participants completed three patient cases in each navigator and were randomly assigned to start with the original or redesigned navigator. Participants were asked to think aloud while performing tasks, and completed a set of surveys after each case and at the end.

Results

Participants were experienced EHR users but were not experienced using either of the two versions of the ambulatory navigator. Time to complete tasks varied widely across participants. Within participants, each of the six patient cases took similar amounts of time, indicating that the cases were of similar complexity. Average time to completion for five of the six patient cases was longer in the new navigator. Scores on the systems usability survey suggested that participants had mixed preferences between the two navigators, with a slight overall preference for the new navigator. Data from the think aloud procedure revealed that all participants encountered usability problems while completing tasks in both navigators and expressed differing preferences between the two navigators.

Discussion

Despite the time and effort given to redesigning the navigator, provider preferences were mixed, and time to complete tasks was longer with the new navigator. This indicates that designing EHR interfaces is a very complex and difficult process, even for a single section of an EHR interface. This study is limited by its small sample size of residents from one specialty at one institution, and participants had varying levels of skill using EHRs. Efforts to redesign EHR interfaces are complicated by differences in user preferences and experiences and tension exists between the benefit of standardization and allowing for customization based on preferences. Future work should further explore how the design process can account for differing preferences in EHR interface design and to measure how changes in EHR interface design impact perceived workload and time to complete tasks.

Acknowledgements

This work was supported by National Science Foundation Award #CMMI-1150057 (JM) and the Agency for Healthcare Research and Quality Award #R01HS022085 (GM)
Figure 1: Old Navigator

Figure 3: Time to Complete Patient Case Comparison

Figure 4: Perceived Workload Comparison

©2015 Epic Systems Corporation. Used with permission.

Figure 2: New Navigator

©2015 Epic Systems Corporation. Used with permission.
Process Mining of Growing Adoption of Genomic Precision Medicine Testing Using Commercial Claims and Encounters Database

Vojtech Huser, MD, PhD

1Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction

Many studies show increasing penetration of genomic testing to support precision medicine. In 2012, the American Medical Association created new codes to facilitate billing of molecular pathology (MoPath) tests using Current Procedural Terminology [CPT] codes 81200 – 81479 and G0452). Many genetic labs admit that such enumeration of genomic tests is incomplete1 and that they often have to resort to generic codes (eg, CPT:81479 “Unlisted molecular pathology procedure”). Despite these limitations, examination of a large administrative dataset can offer unique national level insight into the gradual adoption of precision medicine molecular testing; we present such an analysis. In addition, we reviewed test catalogs of several laboratories to compare genomic tests linked to the same CPT code with a special focus on how exome sequencing is presented and billed by different labs.

Methods

We used MarketScan Commercial Claims and Encounters (CCAE) dataset that contains data from privately insured population and contains patient-level de-identified claims data from inpatient and outpatient visits and pharmacy claims of multiple insurance plans. Access to this dataset was provided by the Innovation in Medical Evidence Development and Surveillance (IMEDS) project of the Reagan-Udall Foundation (RUF) for the Food and Drug Administration. CCAE is one of the largest current healthcare datasets with data on 141.8 million patients (Mean age: 36 (SD:18.57); Time-span: 2003-2013; Drug records: 2.8 billion NDC pharmacy dispensing claims + HCPCS/CPT/ICD9Proc drug codes (inpatient/outpatient claims); Dx:4.6 billion ICD9 codes (inpatient/outpatient claims); Procedures: 10 billion HCPCS/CPT/ICD9Proc codes (inpatient/outpatient claims); Laboratory data and observations: 385 million LOINC-coded outpatient results). We used process mining techniques, Structured Query Language (SQL), R language and tools developed by the Observational Health Data Sciences and Informatics (OHDSI) community.

Results and Discussion

To facilitate comparison and execution on other datasets, we used the R Markdown format designed for reproducible science that combines the analysis code with the resulting analytical report (see project repository for code [*.Rmd] and output reports [*.docx]). Because the CCAE dataset is formatted in the OHDSI Common Data Model (CDM v4; mainly tables PROCEDURE_OCCURENCE, PROCEDURE_COST and PERSON), our analysis can be repeated on refreshed data or on comparable datasets (in CDM format) from other institutions. We extracted 324,334 molecular pathology testing events for a total of 143,811 patients. 62,358 patients (43%) had more than one genomic test. We classified all molecular test claim codes into three categories: (1) gene-specific codes, such as “KRAS gene analysis, codons 12 and 13”; (2) pharmacogenetics codes, such as “CYP2C19 gene analysis, common variants”, and (3) generic codes (CPT Tier 2), such as “Molecular pathology procedure, Level 8 (eg, 26-50 exons)”. Table 1 presents an overview of selected set of parameters for a demonstrative subset of codes. For each test we looked at usage over time (sample shown in Figure 1 for a MLH1 gene test used in Lynch syndrome), associated diagnoses (using CDM’s ‘relevant_condition_concept_id’ column) and average costs (paid by the health plan and the patient) and cost variability. See project repository for a complete list and all generated graphs. This analysis is part of our larger effort to characterize EHR data of genomic precision medicine patients. From an informatics standpoint, our analysis also represents a case study where a distributed cloud database (RedShift engine running on an Amazon cluster of eight dw1.xlarge servers; total of 16 vCPUs, 120GB memory and 16TB storage) provided highly performing analytical environment. The genomic tests extraction that involves a join of two largest tables in the largest IMEDS dataset (10 billion rows with joined cost data) executed in 5.2 minutes.

Claims data show increasing adoption of genomic testing over time. Review of laboratory catalogs reveals significant differences that are also reflected by highly variable cost data. Recent FDA regulatory actions that may limit laboratory-developed tests may lead to further standardization of claims data. Standardization of genomic events using non-CPT based approaches in non-claims-only EHR data can enable much more detailed analyses.
Table 1. Overview of selected genomic tests (using shortened test names)

<table>
<thead>
<tr>
<th>Code</th>
<th>Short Test Name</th>
<th>Count</th>
<th>Avg. Total $ Cost (Patient)</th>
</tr>
</thead>
<tbody>
<tr>
<td>81220</td>
<td>CFTR (eg, cystic fibrosis), gene analysis; common variants</td>
<td>33,496</td>
<td>$658 (340)</td>
</tr>
<tr>
<td>81241</td>
<td>F5, factor V (eg, hereditary hypercoagulability) gene analysis</td>
<td>17,898</td>
<td>$109 (56)</td>
</tr>
<tr>
<td>81240</td>
<td>F2, factor II (eg, hereditary hypercoagulability) gene analysis</td>
<td>16,712</td>
<td>$103 (54)</td>
</tr>
<tr>
<td>81203</td>
<td>MTHFR (eg, hereditary hypercoagulability) gene analysis</td>
<td>14,971</td>
<td>$142 (78)</td>
</tr>
<tr>
<td>81243</td>
<td>FMR1 (eg, fragile X mental retardation) gene analysis</td>
<td>9,601</td>
<td>$145 (63)</td>
</tr>
<tr>
<td>81211</td>
<td>BRCA1, BRCA2 (breast ca), sequence and common variants</td>
<td>6,158</td>
<td>$2,962 (740)</td>
</tr>
<tr>
<td>81213</td>
<td>BRCA1, BRCA2 (breast ca), uncommon variants</td>
<td>4,763</td>
<td>$600 (146)</td>
</tr>
<tr>
<td>81200</td>
<td>ASPA (eg, Caravan disease) gene analysis</td>
<td>4,247</td>
<td>$137 (64)</td>
</tr>
</tbody>
</table>

Gene specific codes (MoPath CPT Tier 1)

<table>
<thead>
<tr>
<th>Code</th>
<th>Short Test Name</th>
<th>Count</th>
<th>Avg. Total $ Cost (Patient)</th>
</tr>
</thead>
<tbody>
<tr>
<td>81401</td>
<td>Molecular pathology proc, Level 2 (eg, 2-10 SNPs)</td>
<td>25,897</td>
<td>$300 (146)</td>
</tr>
<tr>
<td>81479</td>
<td>Unlisted molecular pathology procedure</td>
<td>17,952</td>
<td>$588 (272)</td>
</tr>
<tr>
<td>81400</td>
<td>Molecular pathology proc, Level 1 (eg, 1 SNP)</td>
<td>5,462</td>
<td>$144 (56)</td>
</tr>
<tr>
<td>81403</td>
<td>Molecular pathology proc, Level 4 (single exon)</td>
<td>1,803</td>
<td>$483 (150)</td>
</tr>
<tr>
<td>81404</td>
<td>Molecular pathology proc, Level 5 (eg, 2-5 exons)</td>
<td>1,388</td>
<td>$927 (248)</td>
</tr>
<tr>
<td>81402</td>
<td>Molecular pathology proc, Level 3 (eg, &gt;10 SNPs)</td>
<td>1,024</td>
<td>$273 (103)</td>
</tr>
<tr>
<td>81406</td>
<td>Molecular pathology proc, Level 7 (eg, 11-25 exons)</td>
<td>768</td>
<td>$4,129 (314)</td>
</tr>
<tr>
<td>81405</td>
<td>Molecular pathology proc, Level 6 (eg, 6-10 exons), CYP21A2</td>
<td>549</td>
<td>$1,278 (370)</td>
</tr>
</tbody>
</table>

Pharmacogenomics codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Short Test Name</th>
<th>Count</th>
<th>Avg. Total $ Cost (Patient)</th>
</tr>
</thead>
<tbody>
<tr>
<td>81225</td>
<td>CYP2C19, gene analysis, common variants</td>
<td>9,070</td>
<td>$295 (168)</td>
</tr>
<tr>
<td>81326</td>
<td>CYP2D6, gene analysis, common variants</td>
<td>2,407</td>
<td>$648 (334)</td>
</tr>
<tr>
<td>81227</td>
<td>CYP2C9, gene analysis, common variants</td>
<td>2,220</td>
<td>$406 (182)</td>
</tr>
<tr>
<td>81355</td>
<td>VKORC1 (eg, warfarin metabolism)</td>
<td>1,640</td>
<td>$216 (54)</td>
</tr>
</tbody>
</table>

Figure 1. Graph indicating growing adoption of CPT:81292 test “MLH1 gene (eg, Lynch syndrome), full sequence analysis” by plotting observed test frequency by month

*we plan to repeat the analysis once IMEDS refreshes the CCAE dataset with 2014 data

Acknowledgements: This work was supported by the RUF project IMEDS-SA-0011 and the Intramural Research Program of the NIH, National Library of Medicine. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

References

Project repository (analysis code and full output reports) http://dx.doi.org/10.6084/m9.figshare.1312837
A Road Map for a National Health Information Technology Safety Center

Douglas Johnston, MTS¹, Andrew Gettinger, MD², Kathy Kenyon, JD³, Stephanie Rizk, MS⁴, Colene Byrne, PhD⁵, Linda Dimitropoulos, PhD⁶
¹, ⁴-⁶RTI International, Raleigh-Durham, NC; ²,³Office of the National Coordinator for Health Information Technology, Washington, DC

Abstract
Prior work by federal advisory committees and agencies identified the potential value of a national health IT safety center. We describe the process for developing a road map for such a center – including its vision, objectives, core functions and operating model. Whether or not a center is funded, this road map identifies potential high value functions and activities requiring multi-stakeholder engagement in order to maximize the potential for health IT to make care safer.

Introduction
Recent studies present accumulating evidence that adoption of health IT has made progress toward important quality and safety goals.¹ At the same time, some studies have found that health IT is a contributing factor to adverse events and near misses.²,³ A number of reports have further described health IT safety risks and the actions needed to address them.⁴⁻⁵ Recently, the draft FDASIA report proposed creating a national health IT safety center as a trusted convener of health IT stakeholders in support of these efforts.⁶ In September 2014, the Office of the National Coordinator for Health Information Technology (ONC) contracted with RTI International to convene a multi-stakeholder task force to provide input into a road map for a potential health IT safety center. The methods and inputs used to develop the road map, and its results, are described below.

Methods
To develop this road map, a task force of nationally-recognized experts and stakeholders was convened to provide input on the center’s vision, core functions, and operating model. The task force represented the likely participants and audiences for a national health IT safety center. To guide road map development, the task force was provided a list of operational considerations as well as a summary of federal agency authorities that would limit a center’s abilities to conduct certain activities (Table 1). The task force both defined the general functions of the center, as well as priority areas – such as the development of actionable solutions for high risk safety issues – to ensure value for stakeholders participating in the center. Central to all functions was engaging private-sector stakeholders in ways that support and advance (and do not supplant) their health IT safety initiatives and responsibilities.

The health IT safety center as envisioned would potentially work with individuals and organizations in the private sector and government to create a learning system committed to 1) using health IT to make care safer and 2) to continuously improving the safety of health IT. To meet these fundamental objectives, its activities will lead to improvements in: identification of health IT safety events and hazards; evidence, awareness, and knowledge around health IT safety; health IT safety tools, best practices, interventions, and education; safe use of health IT through improvements in design, usability, competence, and other areas; and uses of health IT to improve patient safety. The center is envisioned as a public-private entity, potentially governed by a multi-stakeholder body of representatives from research and academic institutions, health IT vendors, patient safety organizations, insurers, provider organizations, patient advocates, federal agencies, and others.

Results and Discussion
The health IT safety center road map will be completed in late spring of 2015. In addition to the overview provided above, our proposed presentation in the November 2015 AMIA Annual Symposium will review the core elements of the road map. Our talk will provide AMIA audience members with an opportunity to learn and ask questions about the road map, and to discuss a potential health IT safety center as well as other mechanisms to foster multi-stakeholder collaboration to address the core high value functions and activities identified in the road map.
Table 1. Operational Considerations for a national health IT safety center.

<table>
<thead>
<tr>
<th>Potential center scope of activities</th>
<th>Limitations on center scope of activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Conduct educational programs</td>
<td>• Will not engage in direct investigation or surveillance</td>
</tr>
<tr>
<td>• Promote opportunities for engagement and research</td>
<td>• Will not include operating or funding the operations of a Patient Safety Organization</td>
</tr>
<tr>
<td>• Analyze evidence of health IT-related events and near misses</td>
<td>• Will not include direct data collection</td>
</tr>
<tr>
<td>• Advance tool/intervention development</td>
<td>• Will not include performing functions of Federal Advisory Committees</td>
</tr>
<tr>
<td>• Identify health IT safety goals, priorities, and related measures</td>
<td>• Will not include activities that are exclusively the responsibility of federal entities, and, therefore, cannot be delegated to outside parties, such as the exercise of regulatory authority, establishing government programs, and decision-making related to federal budget expenditures and priorities</td>
</tr>
<tr>
<td>• Support measures and evaluate progress toward goals</td>
<td></td>
</tr>
<tr>
<td>• Collect and share learning/best practices</td>
<td></td>
</tr>
<tr>
<td>• Provide a forum for public and private sector stakeholders</td>
<td></td>
</tr>
</tbody>
</table>

References


A Genome- and Phenome- Wide Study of Diverticulosis

Yoonjung Y. Joo1; Jennifer A. Pacheco1; Loren Armstrong1; William K. Thompson, PhD1; Robert J. Carroll, PhD2; Joshua C. Denny, MD, MS2; Peggy L. Peissig, PhD3; James G. Linneman3; Jyotishman Pathak, PhD4; Girish N Nadkarni, MD, MPH, CPH5; Laura J. Rasmussen-Torvik, PhD, MPH1; M. Geoffrey Hayes, PhD1; Abel N. Kho MD, MS1

1Northwestern University, Chicago, IL; 2Vanderbilt University, Nashville, TN; 3Marshfield Clinic, Marshfield WI; 4Mayo Clinic, Rochester, MN; 5Mount Sinai, New York, NY

Introduction

Genome-wide Association Studies (GWAS) provide hypothesis-free investigation of the whole genome. However, GWAS have limited ability to identify clinically significant variants, and account only for a portion of the predicted phenotypic heritability. In addition, most GWAS-significant single nucleotide polymorphisms (SNPs) are located in intergenic regions, which make the interpretation of biological mechanisms more challenging. As a complement to GWAS, Phenome-wide association studies (PheWAS) offer insights into the biological mechanisms that predict disease susceptibility, and help to determine promising polymorphisms or variants for further study.1

Methods

As part of the eMERGE network,2 we developed and validated a rule-based algorithm to find cases and controls for diverticulosis and diverticulitis in adult eMERGE subjects using electronic health record (EHR) data. At all but two sites, we selected subjects from patients having colonoscopies or abdominal imaging: Natural Language Processing (NLP) was performed on the reports from those procedures, and any subject that had any asserted mention of “diverticul*” in those reports was recorded as cases, plus any with an asserted mention of diverticulitis were noted. Conversely, controls had at least one colonoscopy and no asserted mentions of “diverticul*” in those reports. For two sites where NLP of all reports was not possible, cases were selected as above, plus only those diagnosed (using ICD-9 codes) with diverticulosis or diverticulitis within 7 days after the procedures were selected as cases. Conversely, controls were also the same, but in addition had to have no diagnosis ever for diverticulosis or diverticulitis. Three sites validated3 algorithm performance by chart review of a combined total of 300 randomly selected charts of cases and controls to ensure a high positive predictive value (PPV). We performed GWAS of diverticulosis cases, versus the subset of those with diverticulitis, and versus controls, stratified by race. We selected the top SNPs associated with diverticulosis versus controls, and versus diverticulitis for the PheWAS. We used logistic regression adjusted for age and sex assuming an additive genetic model, on the eMERGE adult cohort of ~38,000 individuals, and also on various subsets, including just those with diverticulosis, and diverticulitis, using the R PheWAS package.4

Results

Validation of the phenotype algorithm resulted in 98-100% PPV at 2 sites for cases and controls, and 88% PPV at a third site for cases, which has subsequently updated their NLP system to correct the errors found. We identified 6297 diverticulosis cases, of which 734 had diverticulitis, and 5090 controls. The GWAS found 39 SNPs to be highly associated (p<10^{-6}) with either diverticulosis and/or diverticulitis and we selected the top SNPs (rs76633992, rs62382461) with the lowest p-values (p=9.28E-08 and p=4.62E-08) that were also near known genes from each population of European and African Americans, respectively, for further investigation by PheWAS. The PheWAS identified several significant associations with cardiac disorders, including abnormal electrocardiogram and cardiac arrhythmia (Figures 1, 2), and with several types of skin disorders (Figure 2).

Discussion and Conclusion

Our findings reinforce the utility of PheWAS as a tool for not only replicating genotype-phenotype associations but also providing insight into potential promising pathways for further investigation. In particular, one of our results replicates the previous GWAS finding of an association between the CAV2 gene and cardiac arrhythmia.5 In addition, as a disease of epithelial cells, the association of these SNPs associated with both diverticulosis and skin disorders is biologically plausible and may suggest a possible disease pathway. Ongoing work is underway to validate the GWAS results for all adult eMERGE diverticulosis case and control subjects, and performing additional PheWAS analyses.
Figure 1. PheWAS plot showing correlation of SNP rs76633992 (correlated with diverticulosis via GWAS) with cardiac arrhythmia in European American cohort.

Figure 2. PheWAS plot showing correlation of SNP rs62382461 (correlated with diverticulosis/diverticulitis via GWAS) with abnormal ECG(electrocardiogram) and various types of skin disorders in African American cohort.

Acknowledgements
The eMERGE Network was initiated and funded by NHGRI through the following grants: U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center).

References
Natural Language Processing facilitates delivery of individualized recommendations at the point of care


Mayo Clinic, Rochester, MN, USA

Introduction

Mayo Clinic has made significant strides in developing knowledge assets and the tools to manage them. To fully realize the value of these assets, it is pertinent not only to embed the knowledge into the clinical workflow but also contextualize the knowledge to individual patients. To ensure the delivery of standard guidelines, we have created care process models (CPMs) for various conditions. However, individualized patient care according to CPMs requires the review and obtainment of relevant clinical information from both structured data fields and unstructured sources of information such as clinical notes, lab reports, etc. While information in structured fields is highly useful, the integrated information from both structured and unstructured repositories at various time points provides comprehensive views during personalized decision-making. Here, we discuss a multi-stage iterative development of an NLP system that extracts data elements needed for automating care process models for hyperlipidemia, atrial fibrillation and heart failure. We also investigated how NLP helps to fill the information gap that exists in structured data sources and empowers clinical decision-making at the point of care.

Methods

As the first step, we developed a gold standard corpus, which was identified through stratified sampling from a retrospective cohort for the data elements pertaining to the three care process models (CPMs) namely Atrial Fibrillation, Heart Failure and Hyperlipidemia. Two nurse abstractors carried out the annotations based on the guidelines (Supplementary File1) outlined by domain experts. The nurse abstractors annotated 165 clinical notes in three batches (55 clinical notes per batch). Based on the gold-standard annotations from clinical notes and standard ontological sources such as ICD9-CM, SNOMED-CT, we developed regular expressions using MedTagger to extract (See Supplementary File2) the data elements from selected sections in clinical notes. The choice of a rule-based approach stems from the overall goal of the project to achieve a very high precision (~100%), at the cost of lower recall. We used an iterative approach (Figure 1) to develop the rule-based NLP system. At the end of the first batch annotation, we used the gold standard annotations to design regular expressions to extract the data elements from clinical notes. We iteratively refined the rules developed to extract data elements based on the analysis carried out after each run.

Results

The system was evaluated against the gold standard annotated data set in an iterative manner. Table1 lists the performance of the NLP system in terms of precision, recall and F-measure (standard metrics for evaluation) against batch2 and batch3 data sets. In the first iteration, the system developed based on batch1 gold standard was evaluated against batch2 data set (Row 2 in Table1). In the second iteration, we refined the rules to fix errors observed in iteration 1 and evaluated the resultant system against the batch3 data set. After a couple of iterations, the system achieved desirable levels of performance. The above evaluation reflects the ability to extract data elements by the NLP algorithm on a smaller data set. We also conducted a pilot evaluation regarding the impact of NLP on individualized care recommendations made on 14,000 patients. Based on the data outlined in Table 2 & Supplementary File3 we infer that when the NLP system decisions when combined with the information in structured data made significant differences to the care recommendations to 1,730 patients. From Table 3, Supplementary File3 and 4 we can infer that the NLP system while reversing the care recommendations for 155 patients, has suggested alternative care recommendations to 1, 885 patients. We will soon realize the robustness of the NLP recommendations as the physicians are reviewing the care recommendations made by MEA in the pilot study.

Conclusions and Future work

We have successfully integrated an NLP system into an EMR agnostic care recommendation solution system, MayoExpertAdvisor. The adopted batch-wise iterative system development has yielded a robust NLP system with acceptable performance.
**Table 1. Evaluation of NLP system on batch2 and batch3 annotation data set**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total data elements annotated</th>
<th>Total data elements extracted</th>
<th>Total Correct</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch2</td>
<td>239</td>
<td>219</td>
<td>206</td>
<td>94.06</td>
<td>86.19</td>
<td>89.96</td>
</tr>
<tr>
<td>Batch3</td>
<td>183</td>
<td>163</td>
<td>160</td>
<td>98.16</td>
<td>87.43</td>
<td>92.49</td>
</tr>
</tbody>
</table>

**Table 2. Assessment of MEA care-recommendations with and without NLP**

<table>
<thead>
<tr>
<th>Care Process Models</th>
<th>Number of patients who received care recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without NLP</td>
</tr>
<tr>
<td>Atrial Fibrillation</td>
<td>753</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>598</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>2951</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>4302</strong></td>
</tr>
</tbody>
</table>

**Table 3. Impact of NLP on patient care-recommendations**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care recommendation reversal</td>
<td>155</td>
</tr>
<tr>
<td>Alternative recommendations by NLP system</td>
<td>1885</td>
</tr>
</tbody>
</table>

**References**

Population Level Clinical Analytics Using the MapReduce Framework and a Production Rule System
Neelima Karipineni, MD, MMSc1,2,3, Howard S. Goldberg, MD1,2,3
1Partners HealthCare System, Wellesley, MA; 2Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA; 3Harvard Medical School, Boston, MA

Introduction

We have demonstrated use of a highly scalable web-service-based decision support service, the Enterprise Clinical Rules Service (ECRS), in several patient-centered clinical decision support applications1-3. We repurposed this system to author population level rules that calculate a risk score for thromboembolic events in patients with non-valvular atrial fibrillation, the CHA2DS2-VASc score4, and determine whether the patient is appropriately anticoagulated. We found utilizing our production web service in batch mode, as we do to retrieve real-time clinical decision support recommendations, to be inefficient for large volume data. Hadoop is an open-source Apache Foundation project which at its core includes an implementation of a parallel computing framework, MapReduce, and a distributed file system, the Hadoop Distributed File System (HDFS). This framework has a number of potential benefits over our current process, including decreased network usage and ease of scaling clusters.

Methods

The ECRS architecture integrates a commercial production rule system, Operational Decision Manager (IBM, Armonk, NY) with internally developed terminology and classifications services. The system allows clinical analysts to write decision support rules in an English-like language against a shared patient information model (PIM), which is populated from input data in standard formats, such as the Continuity of Care Document (CCD). We extracted the rule application as a Java package and used a distribution of the ODM execution engine to execute the rules in the Map portion of a MapReduce process. We tested combinations of custom Map and Reduce implementations, as well as input file formats, on a set of 5,060 PIM XML files. We compared CPU, memory, and disk usage for each, averaging results over three runs. All tests were performed on a single virtual machine running CentOS 6.6 and Hadoop 2.6.0 in pseudo-distributed mode, with 8GB of memory and four Intel(R) Xeon(R) CPU E5-2690 v2 3.00GHz processors, giving us the ability to run three concurrent Mappers per job.

Results

As shown in table 1, all five tested configurations gave us significant performance improvement over sequential calls to a web service, which took approximately 117 minutes with the same data set. As expected, the configuration without a Reduce phase gave us the best performance and least resource utilization. Of the configurations including a Reducer, using the core SequenceFileInputFormat was the most efficient, primarily due to decreased disk reads from HDFS. However, sequence file creation from individual XML files added an additional 1.64 minutes in overhead, negating much of the time performance benefit. The first and second comparisons gave similar performance, with a small, not statistically significant advantage in the configuration deserializing the Patient object from XML in the Mapper rather than in the record reader.

Discussion

While our distributed process reduced runtime by as much as 97%, we found a fundamental mismatch between our use case and the MapReduce framework: the framework is optimized for simple logical operations performed on a very large amount of unstructured or semi-structured data provided in a few large files. We, however, tested an implementation of fairly complex logic on a small amount of well-structured data provided in many small files. The process of extracting data from clinical systems, transforming it into interoperable models, and running classification and translation terminology services far outweighed the time taken to execute the rules by any compared method. Using a MapReduce framework would be best suited to persistent, pre-existing patient data sets. On an average day, the Partners HealthCare clinical document repository stores upwards of 12,000 CCDs for outpatient clinic visits alone. In the future, we would explore techniques for de-identification and institutional barriers to use of existing patient data repositories such as this for large scale analyses. We would further explore scalability in terms of data set size and cluster size.
 Configuration (InputFormat Mapper <K,V> Reducer <K,V>) | Elapsed time (min) | Total CPU vcore-seconds | Total memory megabyte-seconds | Total local data (bytes) read written |
--- | --- | --- | --- | --- |
4. CombineXMLFileInputFormat <XML Filename, Patient XML> No Reducer | 3.38 | 482340 | 493916160 | 0 319880 |
5. SequenceFileInputFormat <XML Filename, Patient XML> <PatientID, Recommendation Object> | 4.91 | 834730 | 854763520 | 25296805 51021590 |

Table 1: MapReduce algorithm implementations and associated resource utilization
<K, V> indicates Key and Value object types of input parameters. Local data read/written does not include HDFS.

References

Demonstrating the Advantages of Applying Data Mining Techniques on Time-Dependent Electronic Medical Records

Uri Kartoun, PhD1,2, Vishesh Kumar, MD1,2, Su-Chun Cheng, ScD3, Sheng Yu, PhD3, Katherine Liao, MD, MPH2,4, Elizabeth Karlson, MD2,4, Ashwin Ananthakrishnan, MBBS, MPH2,5, Zongqi Xia, MD, PhD2,5, Vivian Gainer, MS6, Andrew Cagan, BSc6, Guergana Savova, PhD2,7, Pei Chen, MS2,7, Shawn Murphy, MD, PhD6, Susanne Churchill, PhD5, Isaac Kohane, MD, PhD2,5, Peter Szolovits, PhD5, Tianxi Cai, ScD3, Stanley Y. Shaw, MD, PhD1,2

1. Center for Systems Biology, Massachusetts General Hospital (MGH), Boston, MA; 2. Harvard Medical School, Boston, MA; 3. Department of Biostatistics, Harvard School of Public Health, Boston, MA; 4. Division of Rheumatology, Brigham and Women’s Hospital (BWH), Boston, MA; 5. i2b2 National Center for Biomedical Computing, BWH, Boston, MA; 6. Research Computing, MGH, Boston, MA; 7. Clinical Natural Language Processing Program, Boston Children’s Hospital, Boston, MA; 8. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

**Summary:** We demonstrate several advantages of applying data mining techniques on time-dependent Electronic Medical Records (EMR), specifically: 1) combining structured and unstructured variables improves the accuracy of a type-2 diabetes (T2D) classification algorithm, 2) conducting a quantitative survey of multiple comorbidities is important in T2D especially cardiovascular complications with hazard ratios, 3) analyzing time dependent variables can clarify time dependent contributions to variety of comorbidities, and specifically of the “obesity paradox”, and 4) demonstrating that an unbiased examination of physician treatment patterns reveals changes over time consistent with clinical trials.

**Background:** Cohorts assembled from EMR present a potentially powerful resource to study T2D and cardiovascular complications at population scale. Recent reports have demonstrated the utility of EMR analysis to discover genotype-phenotype correlations, sub-categories of disease, and adverse drug events.

**Methods:** We developed a classification algorithm to identify T2D patients based on characteristics including clinical notes, diagnosis and procedure codes, medications, and laboratory tests. We analyzed an EMR database at MGH and BWH considering patients who received care between 1990 - 2013. We applied logistic regression with the adaptive LASSO using different combinations of variables such as structured variables only, unstructured variables only, and combination of all variables. To determine the level of association between clinical and demographic variables with mortality we developed baseline and lagged-time varying Cox regression models that included an adjustment to ethnicity and time varying covariates. To assess how therapeutic choices change over time, we calculated sparse covariance matrices for heart failure related concepts extracted from clinical notes.

**Results:** Our classification algorithm identified 65,099 T2D patients with a specificity of 97% and PPV of 96% based on “gold standard” physician chart review. 56,691 patients (87.1%) had two and 38,449 patients (59.1%) had four or more chronic conditions, demonstrating the complexity of the cohort. Cox regression models indicated statistically significant HRs > 1 for CHF, CAD, and CVD, and HRs < 1 for PCI and CABG. Increasing BMI was associated with lower mortality as compared to the reference BMI (< 25 kg/m²). Further stratifying the results into 1, 3 and 5 years analysis, this “obesity paradox” is strikingly obvious at short-term follow-up of 1 year, suggesting that patients with low BMI were suffering from chronic medical conditions (e.g., malignancy or inflammatory conditions) increasing their 1 year mortality. However, at 3 and 5 years follow-up, we do see increase in mortality with increasing BMI levels likely related to increase in the burden of cardiovascular events.

**Discussion:** We implemented classification, prediction, and natural language processing techniques in multiple scenarios to create and to analyze a highly complex and large cohort. This cohort recapitulates many findings from traditionally ascertained cohorts while enabling additional analyses (e.g., utilizing physician notes or richer temporal data), illustrating its utility for a variety of discovery efforts.
Table 1. 65,099-patient cohort characteristics considering diagnosis codes

(a) Most common comorbidities and combinations of comorbidities (for no comorbidities: n = 1,117):

<table>
<thead>
<tr>
<th>Most Common</th>
<th>Diabetes (n = 5,422)</th>
<th>Exactly 1 comorbidity</th>
<th>Hypertension (n = 7,152)</th>
<th>Exactly 2 comorbidities</th>
<th>Hyperlipidemia (n = 11,090)</th>
<th>Exactly 3 comorbidities</th>
<th>Diabetes &amp; Hypertension &amp; Hyperlipidemia (5,795 patients) (n = 10,643)</th>
<th>Exactly 4 comorbidities</th>
<th>Diabetes &amp; Hypertension &amp; Hyperlipidemia &amp; Arthritis &amp; Depression (2,421 patients) (n = 8,664)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>78.0%</td>
<td>85.6%</td>
<td>92.0%</td>
<td>94.1%</td>
<td>94.8%</td>
<td>95.6%</td>
<td>Diabetes &amp; Hypertension &amp; Hyperlipidemia (5795 patients)</td>
<td>96.1%</td>
<td>Diabetes &amp; Hypertension &amp; Hyperlipidemia &amp; Arthritis &amp; Depression (571 patients)</td>
</tr>
<tr>
<td>Most Common</td>
<td>Hypertension</td>
<td>6.0%</td>
<td>48.8%</td>
<td>81.2%</td>
<td>88.2%</td>
<td>91.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hyperlipidemia</td>
<td>3.4%</td>
<td>27.2%</td>
<td>65.6%</td>
<td>73.9%</td>
<td>76.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Most common comorbidities considering a minimum number of comorbidities threshold per patient:

<table>
<thead>
<tr>
<th>Comorbidities</th>
<th>≥ 0 Comorbidities (n = 63,230)</th>
<th>≥ 1 Comorbidities (n = 62,113)</th>
<th>≥ 2 Comorbidities (n = 56,691)</th>
<th>≥ 3 Comorbidities (n = 49,539)</th>
<th>≥ 4 Comorbidities (n = 38,449)</th>
<th>≥ 5 Comorbidities (n = 27,806)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>90.6%</td>
<td>92.2%</td>
<td>93.6%</td>
<td>94.8%</td>
<td>95.6%</td>
<td>96.1%</td>
</tr>
<tr>
<td>Hypertension</td>
<td>76.7%</td>
<td>78.1%</td>
<td>85.0%</td>
<td>90.2%</td>
<td>92.8%</td>
<td>94.6%</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>64.0%</td>
<td>65.2%</td>
<td>71.1%</td>
<td>77.4%</td>
<td>80.8%</td>
<td>83.5%</td>
</tr>
</tbody>
</table>

(c) Prevalence of additional common combinations of comorbidities:

<table>
<thead>
<tr>
<th>Combination</th>
<th>Patients associated with the combination (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes &amp; Hypertension</td>
<td>69.7%</td>
</tr>
<tr>
<td>Diabetes &amp; Hyperlipidemia</td>
<td>58.5%</td>
</tr>
<tr>
<td>Hypertension &amp; Hyperlipidemia</td>
<td>55.7%</td>
</tr>
<tr>
<td>Ischemic Heart Disease &amp; Hypertension</td>
<td>18.6%</td>
</tr>
<tr>
<td>Ischemic Heart Disease &amp; Hyperlipidemia</td>
<td>17.1%</td>
</tr>
</tbody>
</table>
Using Patient-Centered Technological Design to Improve Inpatient Fall Prevention

Zachary Katsulis¹, Waiyin Leung, MS², Awatef Ergai, PhD¹, Laura Schenkel¹, Amisha Rai, PA-C⁴, Jason Adelman, MD⁴, James Benneyan, PhD¹, David Bates, MD²,³, Patricia C. Dykes, PhD RN²,³

¹Healthcare Systems Engineering Institute at Northeastern University, Boston, MA; ²Brigham and Women’s Hospital, Boston, MA; ³Harvard Medical School, Boston, MA; ⁴Montefiore Medical Center, Bronx, NY

Introduction: Falls are a serious patient safety issue. Each year, somewhere between 700,000 and 1,000,000 people in the United States fall in the hospital.¹ Hospitalized patients are especially at increased risks for falls and fall injuries.² Most of these falls, however, are considered “preventable” by the Centers for Medicare & Medicaid Services (CMS).³ An earlier study we conducted tested a FALL TIPS (Tailoring Interventions for Patient Safety) Toolkit that used HIT to link routine nursing fall risk with evidence-based interventions to prevent patient falls. The study demonstrated that fall rates were reduced by 22% in intervention units and the toolkit was particularly effective with patients aged 65 years and older.⁴ Based on previous findings, an extension project funded by the Agency for Healthcare Research and Quality (AHRQ), called the Patient-Centered Fall Prevention Toolkit, seeks to use a patient-centered approach to improve inpatient fall prevention through the use of existing iPads at the patient’s bedside. Keeping human factors analysis in mind, patient surveys were conducted at two hospitals serving different socio-economic populations to understand how demographic factors might affect patients’ openness or perceived ability to use technology at the bedside. Using these results we aim to develop an initial patient-centered electronic prototype that accommodates all users in achieving the safest care.

Methods: We conducted patient surveys in pilot units participating on our Patient Safety Learning Labs project (e.g. oncology units at Brigham and Women’s Hospital (BWH) and medicine/surgical units at Montefiore Medical Center (MMC)). Three of the survey questions, seen in Figure 1, related to the patient’s openness, ability, and use of technology. A five point Likert scale was used, with increasing agreement or hours of use corresponding to the increasing scale. Demographic information including age, ethnicity, and highest level of education completed were collected. We compared user preference by analyzing the difference in responses between BWH and MMC, as well as between older (65 or older) and younger patients (less than 65). Due to the inability to assume normal distribution, the comparison of response means between older and younger patients was conducted using a two tailed T-test; whereas a two tailed Z-test was used to compare response means between BWH and MMC (N>30 for each group). Calculated p-values were used to determine significant difference (p-value<0.05) between means.

Results: Table 1 presents the three technology-related patient survey questions along with their calculated p-value and corresponding group means at MMC and BWH. Of these three survey questions, the two patient age groups showed significant difference (p-value = 0.0117 < α = 0.05) for only question 11 at BWH. Figure 1 shows the disparity in demographics between the patient populations at the two hospitals. In addition to demographics, roughly 46% of patients at MMC received some form of higher education (some college or greater) in comparison to 76% of patients at BWH. Table 2 showcases there was no significant difference between patient openness, ability, and average use of technology despite the disparity in ethnicity and education level at the two hospitals.

Discussion: At BWH, older patients reported that they are more likely to be less open to using an electronic device in their care plan than younger patients. This same trend was seen at MMC but was not significant. We believe this to be due to the small sample of older patients and we would expect to see the same trend given a greater sample size. We believe it may be beneficial to consider involving family members in using HIT for fall prevention particularly for frail senior patients, as caregivers (often younger and not functionally impaired) can play an important role but are often overlooked targets for expanding portal use at the bedside.⁵ Based on our data, we believe that on our target units, ethnicity and education level will have no significant impact on the openness, ability, or average use of an electronic device, such as an iPad, in their care plan. This supports prior research that blacks and whites have nearly identical rates of smartphone ownership and use, despite a previous “black/white ‘digital divide’”⁶ Using these survey results and patient feedback, we will continuously engage patients in the iterative design process of our iPad application with the ultimate goal of increasing usability and decreasing inpatient falls.
Table 1. Patient survey questions 11-13 along with each age group’s mean and the calculated p-value.

<table>
<thead>
<tr>
<th>Survey Question</th>
<th>MMC (N=32)</th>
<th>BWH (N=54)</th>
<th>p-value</th>
<th>MMC (N=32)</th>
<th>BWH (N=54)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. I am open to using an electronic device with my provider to learn about fall prevention.</td>
<td>3.429</td>
<td>3.541</td>
<td>0.360</td>
<td>2.800</td>
<td>2.688</td>
<td>0.012</td>
</tr>
<tr>
<td>12. I am able to use an iPad without the assistance of a family member.</td>
<td>4.000</td>
<td>3.757</td>
<td>0.354</td>
<td>3.200</td>
<td>3.118</td>
<td>0.097</td>
</tr>
<tr>
<td>13. How many hours per day on average do you use an electronic device such as a computer or an iPad?</td>
<td>3.667</td>
<td>3.649</td>
<td>0.113</td>
<td>2.286</td>
<td>3.059</td>
<td>0.109</td>
</tr>
</tbody>
</table>

Figure 1. Pie charts displaying demographic disparity between the two hospital patient populations.

Table 2. Patient survey questions 11-13 along with calculated mean, variance, and p-values at the two hospitals.

<table>
<thead>
<tr>
<th>Survey Question</th>
<th>MMC (N=32)</th>
<th>BWH (N=54)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. I am open to using an electronic device with my provider to learn about fall prevention.</td>
<td>Mean: 3.367</td>
<td>Mean: 3.283</td>
<td>0.766</td>
</tr>
<tr>
<td></td>
<td>Variance: 1.620</td>
<td>Variance: 1.322</td>
<td></td>
</tr>
<tr>
<td>12. I am able to use an iPad without the assistance of a family member.</td>
<td>Mean: 3.828</td>
<td>Mean: 3.585</td>
<td>0.423</td>
</tr>
<tr>
<td></td>
<td>Variance: 1.719</td>
<td>Variance: 1.709</td>
<td></td>
</tr>
<tr>
<td>13. How many hours per day on average do you use an electronic device such as a computer or an iPad?</td>
<td>Mean: 3.438</td>
<td>Mean: 3.463</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>Variance: 2.641</td>
<td>Variance: 1.574</td>
<td></td>
</tr>
</tbody>
</table>

References:
Identifying and Understanding Data Quality Issues in a Pediatric Distributed Research Network

Ritu Khare, PhD\textsuperscript{a}, Levon H. Utidjian, MD\textsuperscript{a}, Greg Schulte, MS\textsuperscript{a}, Keith Marsolo, PhD\textsuperscript{c}, L. Charles Bailey, MD, PhD\textsuperscript{a}

\textsuperscript{a}Children’s Hospital of Philadelphia, Philadelphia, PA, \textsuperscript{b}Children’s Hospital Colorado, Aurora, CO, \textsuperscript{c}Cincinnati Children’s Hospital Medical Center, Cincinnati, OH

Introduction
Collaborations across multiple institutions are very essential to achieve adequate cohort sizes in pediatrics research\textsuperscript{1}. PEDSnet is a newly established clinical data research network (CDRN) that aggregates electronic health record (EHR) data from eight of the nation’s largest children’s hospitals\textsuperscript{2,3}. With the ultimate goal of supporting a variety of comparative effectiveness research, a prerequisite in PEDSnet is to ensure that the network’s data is “high quality.” The prominent challenges include the lack of EHR data’s fitness for immediate research use, semantic heterogeneity across systems, and data peculiarities in pediatrics\textsuperscript{1,4}. While previous studies have presented frameworks and techniques to validate the EHR-derived data, the process of data quality assessment in CDRNs continues to remain “behind the scenes” with no published empirical results\textsuperscript{4,5,6}. In this study, we implement a comprehensive set of validity checks in PEDSnet to identify, understand, and report a range of data quality issues (see Table 1)\textsuperscript{4,6}.

Methods
The PEDSnet network uses the OMOP Common Data Model (CDM), a widely accepted schema for observational medical data\textsuperscript{7}. Each partner site prepared an instance of the CDM by performing the extract-transform-load (ETL) operations from its EHR according to network-wide conventions. In this study, we focused on attribute-level data quality assessment\textsuperscript{5}, and developed data analysis scripts to ensure adherence to the CDM, perform data domain checks, and compute frequency distributions\textsuperscript{6}. The output report comprises a visual summary (e.g. bar graphs) and a list of automatically detected data quality issues (e.g. out-of-range values) for each attribute. We executed the scripts on each site’s data, and reviewed the graphs to identify additional issues (e.g. unusual shape or peaks). Next, we classified each issue as an “ETL issue” that could be resolved by fixing the ETL logic or a “provenance issue” that exists due to an anomaly, data characteristic, or an error in EHR\textsuperscript{9}. Finally, each issue was communicated to the originating site, which was responsible for validating the cause of the issue and resolving the issue.

Results
At the current stage of this project, we have collected data from all partner sites (total 8), representing 4.4 million children and over 75 million encounters. Table 2 organizes the total number of issues by various data quality dimensions. The fidelity dimension corresponded to the cases where the distribution of the EHR values did not match with that in the CDM, due to ETL errors or due to data characteristics, such as differences in granularities between the coding systems used in the source (e.g. ICD-9) and the CDM (e.g. SNOMED CT). The consistency dimension included the cases where the attribute values were not aligned with the conventions (ETL issue), or were abnormal, e.g. a gestational age value greater than 42 weeks (provenance issue). The accuracy dimension included discrepancies between sites, e.g. significant differences in variation of body weights. Finally, the feasibility dimension corresponded to missing data in the EHR or incomplete ETL mappings.

Discussion
A key challenge in building a CDRN is to define and achieve an optimal degree of data quality. In this study, we have conducted a data quality assessment of the PEDSnet network using rigorous data checks and manual reviews of statistical plots. While this study only focused on attribute level analysis, we have learned several important lessons. In spite of defining network-wide conventions, hosting a shared repository of ETL scripts\textsuperscript{8}, and organizing regular web conferences across sites, we identified over 320 data quality issues across three sites. This strongly suggests that proactive project management and documentation are not sufficient to ensure data validity in a CDRN. We judged nearly 37% of the identified issues as ETL mistakes, and despite the diversity of experiences and backgrounds of various teams, the site-wise contributions to these issues were significantly similar. This reinforces that a formal data quality assessment process is critical in building a CDRN. Our future work includes operationalizing advanced assessment levels\textsuperscript{5}, extending to more clinical domains, and reporting longitudinal results.

Funding: This work was supported by PCORI Contract CDRN-1306-01556.
Table 1. Data Quality Dimensions in PEDSnet

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Definition</th>
<th>Example Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fidelity (i.e. processing integrity)</td>
<td>the degree to which PEDSnet data correctly reflects source systems data</td>
<td>The distributions of gender values do not match between the source system (EHR) and the derived PEDSnet secondary dataset for a given site.</td>
</tr>
<tr>
<td>Consistency (i.e. internal validity)</td>
<td>the degree to which a specific type of information is recorded in the same way in the different data sources contributing to PEDSnet data</td>
<td>A patient with a recorded weight observation of 40,229 kg, or a patient having over 4,000 encounters</td>
</tr>
<tr>
<td>Accuracy (i.e. external validity)</td>
<td>the degree to which PEDSnet data accurately reflects the clinical characteristics of patients</td>
<td>A site having significantly larger ratio for average number of observations per patient</td>
</tr>
<tr>
<td>Completeness (i.e. feasibility)</td>
<td>the degree to which a given type of information is actually collected and available in PEDSnet</td>
<td>The recorded gestational age is missing for 70% of the patients at a given site</td>
</tr>
</tbody>
</table>

Table 2. Total Number (i.e. Counts) of Data Quality Issues in PEDSnet Across 8 Sites

<table>
<thead>
<tr>
<th></th>
<th>Fidelity</th>
<th>Consistency</th>
<th>Accuracy</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETL issue</td>
<td>12</td>
<td>48</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>Provenance issue</td>
<td>7</td>
<td>123</td>
<td>28</td>
<td>48</td>
</tr>
</tbody>
</table>

References

Evaluating the Effects of Cognitive Support on Interpreting ICU Patient Data

Peter Killoran, MD^1,2; Swaroop Gantela, MD^1; Sahiti Myneni, PhD^1; Khalid Almoosa, MD^2; Bela Patel, MD^2; Thomas G. Kannampallil, MS^3; Vimla L. Patel, PhD^4; Trevor Cohen, MBChB, PhD^1.

^1School of Biomedical Informatics and ^2School of Medicine, The University of Texas Health Science Center at Houston, Houston, TX. ^3College of Medicine. University of Illinois at Chicago, Chicago, IL; ^4New York Academy of Medicine. New York, NY.

**Introduction**: Clinicians’ attention is a precious resource, often consumed by the need to aggregate and synthesize information while making clinical decisions^1^. Cognitive Support Systems (CSS) seek to mitigate these cognitive burdens while also enhancing the clinical comprehension necessary to make decisions^2^.

In this work, we characterize the effects of a CSS (SIRSi) designed to organize ICU data in accordance with frameworks of expert clinical decision making regarding Systemic Inflammatory Response Syndrome (SIRS). SIRSi organizes relevant information along “intermediate constructs” – diagnostically and prognostically relevant clusters of clinical information that are representative of experts’ knowledge organization^3^.

In this study, we describe the effects of SIRSi on clinical comprehension based on the analysis of verbal (“think-aloud”) protocols produced by 19 participants (3 Attending and 16 Resident physicians).

**Methods**: SIRSi’s architecture includes web application and knowledge engine servers for executing semantic queries using a custom ontology to retrieve data from MIMIC-2, a de-identified ICU database^5^ (see ^3^ for further details). Our experimental design involved creation of two different interfaces for viewing the same data, one organized by “intermediate constructs” (Concept), and the other organized by data source, as is typical in most contemporary Electronic Health Record systems (Control) (Figure 1). Two patient cases of similar acuity were viewed by each subject after reading a short vignette summarizing the patient’s clinical condition. Assignment of the case to the interface (Concept vs. Control) and the viewing order were randomized for each subject. Subjects then completed a think aloud protocol (in the tradition of ^6^) as they viewed patient data. Participant verbalizations were transcribed, coded, and both the order and frequency of data element mentions were extracted. While data were normalized for all data elements within a given subject, our analysis focused on how subjects attended to data concerning the criteria for SIRS (PaCO2, Respiratory Rate, Heart Rate, White Blood Cell Count, and Temperature). Of note, criteria for SIRS were objectively met for both patient cases.

A transition probability matrix, representing the sequences of transition across data elements was also generated.

**Results**: As expected, the order in which each data element was considered closely followed the interface layout. Participants generally followed a linear, top-down viewing order for both the Concept and Control views (Figure 2). However, attention to the SIRS criteria, a critical component of SIRS-related decision making, was greater for 18/19 participants when using the Concept view. Comparison of the frequency with which SIRS-related variables were mentioned across all participants also showed significantly more attention to SIRS criteria when using the Concept vs. the Control interface (paired t-test, t(18)=4.12, p<0.001). Of note, six out of the nineteen participants using the Control interface did not mention the PaCO2, and two of these did not mention respiratory rate, indicating that they had not adequately assessed the respiratory status of the patient. All six of these participants showed better performance using the Concept interface, as suggested by greater attention to the SIRS criteria, with only one participant failing to mention respiratory rate when using the Concept view.

**Discussion**: Participants generally attended to information in the order in which it was presented. These results are consistent with previous work showing that the layout of information in a clinical display has a significant impact on its interpretation^1^.

Furthermore, our findings show significantly greater attention to SIRS-related data variables with cognitive support. These findings suggest that in the context of complex clinical data associated with critically ill patients, CSS offer advantages over traditional information displays. Future work is needed to validate these findings in real clinical settings.
Figure 1. SIRSi web application interfaces (left: Concept; right: Control)

Figure 2. Transition counts between elements ordered by appearance on screen (left: Concept; right: Control).

<table>
<thead>
<tr>
<th>Concept</th>
<th>PaCO2</th>
<th>HR</th>
<th>RR</th>
<th>Temp</th>
<th>WBC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.032</td>
<td>0.051</td>
<td>0.038</td>
<td>0.028</td>
<td>0.025</td>
<td>0.174</td>
</tr>
<tr>
<td>STD</td>
<td>0.011</td>
<td>0.031</td>
<td>0.020</td>
<td>0.021</td>
<td>0.011</td>
<td>0.059</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Control</th>
<th>PaCO2</th>
<th>HR</th>
<th>RR</th>
<th>Temp</th>
<th>WBC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.026</td>
<td>0.029</td>
<td>0.027</td>
<td>0.025</td>
<td>0.024</td>
<td>0.111</td>
</tr>
<tr>
<td>STD</td>
<td>0.011</td>
<td>0.015</td>
<td>0.009</td>
<td>0.009</td>
<td>0.007</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Table 1. Average normalized count of SIRS related concepts for each interface.

This project was supported by Grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects (SHARP) from the Office of the National Coordinator for Health Information Technology. We acknowledge Zhe Li, Ram Vedam and Manuel Wahle for SIRSi development work.

References

A Framework for Person-centered, Community-wide Care Coordination

Katherine K. Kim, PhD, MPH, MBA1, Janice F. Bell, PhD, MPH, MN1, Charles Boicey, MS, RN-BC, CPHIMS2, Janet Freeman-Daily3, Susan Hull, MSN, RN-4, Anna McCollister-Slipp5, Jill G. Joseph, MD, PhD1

1University of California Davis, Sacramento, CA; 2State University of New York Stony Brook, Stony Brook, NY; 3Seattle, WA; 4Wellspring Consulting, Healdsburg, CA; 5Washington, D.C.

Introduction and Objective

The increasing prevalence of complex chronic conditions such as diabetes, heart disease, and cancer is of great concern. Health care is fragmented and involves diverse participants and resources, pointing to the need for community-wide care coordination (CWCC). While technology may help, little is known about the specific coordination challenges for individuals or about the dynamic nature of coordination across teams and timespans.

Materials and Methods

Content analysis was applied to publicly available data from a one-hour #HCLDR tweetchat held on 10/28/2014. Data were collected from three Twitter search engines to assure complete data capture. 1569 tweets and retweets were posted by 134 participants. 108 tweets pertained to the question “What are the most important challenges that patients, family members, caregivers face in coordinating care?” A framework was developed from literature review1,2 and content analysis. The study team included patients, clinicians, researchers, and informaticists.

Results

The five most frequently mentioned themes are related to patient, family, caregiver challenges and needs in care coordination include: layers and silos of care (25%), data sharing (17%), verbal communication (13%), shared care planning (10%), and family, caregiver needs, expectations, burdens (9%). These data highlighted requirements for HIT to address multiplicity of relationships, dynamic vs. static states, and comprehensive vs. compartmentalized information. Examples of HIT functions for these needs, mapped to themes are suggested (Table 1). The CWCC conceptual framework offers a way to organize HIT functions to assure these needs are addressed. Domains 1-3 are principles while 4-6 indicate dynamic states requiring coordination.

1. Person-centered Coordination: Empower individuals to exercise autonomy, collaborate in decision making, and optimize coordination. Support teams in delivering coordination activities that respond to individuals’ values, needs, and preferences.

2. Shared Care Planning: An inclusive (of individual, family teams, healthcare teams, and/or community teams as appropriate) process of comprehensive assessment, goal-setting and planning, implementation, and evaluation of an individuals’ course of health over the lifespan. The resulting documentation, a shared care plan, may have greater detail on shorter time periods when intensive focus is necessary.

3. Health Information Technology Enablement: Help individuals and teams fulfill CWCC activities with the information and tools to achieve individuals’ health goals, efficiently manage groups of patients, contribute to population health goals. Enable coordination at points of need.

4. Within teams: Three types of teams are family teams, healthcare teams, and community teams. Within each team there are certain roles and responsibilities, specific activities that regularly occur, particular information that is helpful, and unique workflows.

5. Across teams: The individual, family teams, healthcare teams, and community teams interact with each other. The interactions among teams vary in depth and breadth.

6. Over time: CWCC perspective on time is over the continuum of care, and the lifespan of a person. Different teams are active at different times, and the level of participation also varies.

Discussion and Conclusion

Discussions about the need for CWCC must enable dynamic collaboration, be flexible to support changing informational needs, and adapt to multi-setting, multi-team workflows. Mobile applications and social networking technologies can be used to tailor data collection and display while supporting user-driven communication. These applications should be integrated with electronic health records and personal health information tools.
<table>
<thead>
<tr>
<th>Challenge/ Illustrative Tweets</th>
<th>HIT Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Layers and Silos of Care</strong></td>
<td>Navigational resources for planning and managing logistics</td>
</tr>
<tr>
<td>“Each additional provider adds another layer of coordination for families.”</td>
<td>Shared care planning process that offers opportunity for input by all teams (see shared care planning below)</td>
</tr>
<tr>
<td>“When disease not clearly within one specialty, hard to coordinate across disciplines; diseases don't always fit boxes we make!”</td>
<td>Multi-directional health information exchange for timely health data shared among individuals and healthcare teams that are actively involved in treatment and care coordination</td>
</tr>
<tr>
<td><strong>Data Sharing</strong></td>
<td>Secure, standards based ad hoc transfer of data, e.g. DIRECT encrypted email</td>
</tr>
<tr>
<td>“chronic/serious illness generates a ton of healthcare records--my file cabinet is 3 ft deep with</td>
<td>Longitudinal record or personal health record (PHR) owned by patient</td>
</tr>
<tr>
<td>“our #HIE exchange is provider centric not person centric. Where is my longitudinal record in a trusted place easily accessible?”</td>
<td>Dynamic, portable consent for preferences about data sharing and care coordination activities that apply to all the HIT functions in this table.</td>
</tr>
<tr>
<td><strong>Verbal Communication</strong></td>
<td>Patient-focused educational resources and decision-making aids, accessible in terms of language, literacy and meaningfulness, targeted to the phase of care and services being provided, and offered close in time to those needs</td>
</tr>
<tr>
<td>“making sure everyone understands terms, next steps, and actually hears the conversation.”</td>
<td>Audio/video case conferences involving patients and other teams as needed</td>
</tr>
<tr>
<td>“two family members can talk to same doctor and come away with different impression of issue/severity”</td>
<td>Follow-up communications and questions through convenient modes such as email and videoconference, based on patient preference, particularly during transitions</td>
</tr>
<tr>
<td>“Coordination of care requires us to speak the same language both literally and figuratively”</td>
<td></td>
</tr>
<tr>
<td><strong>Shared Care Planning</strong></td>
<td>Access to a longitudinal shared care plan addressing the patients’ health goals as well as clinical goals, across the continuum of care and lifespan, planned actions, schedules, interdependencies between teams, contacts, responsibilities and decision-making authority</td>
</tr>
<tr>
<td>“Getting other healthcare institutions to cooperate with patient’s care coordinators”</td>
<td>Coordinated shared care plan implementation that reflects patient preferences and priorities as well as input from all teams</td>
</tr>
<tr>
<td>“Deciding to what extent we will go to achieve a treatment goal. May vary from patient to family to provider”</td>
<td>Workflow management and coordination support across teams</td>
</tr>
<tr>
<td></td>
<td>Scheduling tools across diverse teams and settings</td>
</tr>
<tr>
<td></td>
<td>Transparent monitoring of progress against goals</td>
</tr>
<tr>
<td></td>
<td>Ability to refer to/communicate with community teams</td>
</tr>
<tr>
<td><strong>Family/Caregiver Needs, Expectations, Burden</strong></td>
<td>Shared care planning process that assesses whether family/caregivers are available and if so, their level of caregiving and care coordination capacity</td>
</tr>
<tr>
<td>“Coordinating 45+ peeps. See Alexis' care map, exhausted just to draw it, living it brings tears freq #hcldr</td>
<td>Participation of family in shared care planning and access to shared care plan</td>
</tr>
<tr>
<td><a href="http://t.co/EkYUFrdqby%E2%80%9D">http://t.co/EkYUFrdqby”</a></td>
<td>Easy and convenient modes of communication with community services</td>
</tr>
<tr>
<td>“Expectation that the patient has an easily available and willing family member to coordinate care is unfair... often not so.”</td>
<td></td>
</tr>
</tbody>
</table>

References

Representing Nursing Content within a Multi-Disciplinary Terminology

Tae Youn Kim, PhD, RN1,2, Nicholas Hardiker, PhD, RN2,3, Penni Hernandez, ND, RN4, Jane Millar4

1University of California Davis, Sacramento, CA, USA
2International Council of Nurses, Geneva, Switzerland
3University of Salford, Salford, UK
4International Health Terminology Standards Development Organisation, Copenhagen, Denmark

Introduction
The International Classification for Nursing Practice (ICNP) is a logic-based nursing terminology developed by the International Council of Nurses (ICN) to support the documentation and exchange of nursing data worldwide.1,2 As SNOMED® CT is considered to be one of the most comprehensive multi-disciplinary terminologies for healthcare, the representation of nursing domain knowledge within SNOMED CT is critical if the terminology is to capture a holistic view of healthcare. Under a collaboration agreement between ICN and the International Health Terminology Standards Development Organisation (IHTSDO),3 the authors agreed work to build linkages for nursing diagnostic/outcome concepts, working with the IHTSDO Nursing Special Interest Group (SIG) to ensure broader involvement of the profession. The objective of this work was to facilitate a transformation pathway between ICNP-encoded data and SNOMED CT, to ensure that nursing continues to be reflected in global multidisciplinary healthcare information systems. This involved a gap analysis and production of an equivalence table.

Methods
A series of activities were undertaken to identify and validate equivalences within SNOMED CT for the 783 pre-coordinated diagnosis or outcome statements within the 2013 release of ICNP, drawing together the results of previous work: 1) pre-existing equivalencies4 within the Unified Medical Language System (UMLS) maintained by the U.S. National Library Medicine5; and 2) candidate equivalencies identified manually for the remaining statements.6 The combined equivalency table was then validated by the IHTSDO Nursing SIG. Any disagreements were further discussed with the ICN eHealth team and IHTSDO terminologists until consensus was reached. Work began in 2011 and has continued through various releases of both terminologies.

Results
The first two steps yielded 196 equivalencies within UMLS and 175 from the manually identified equivalencies. Following validation, there were a total of 336 equivalencies (43% of the total number of ICNP statements). Certain nursing diagnosis/outcome statements within ICNP are negative in nature (relating to illness) (n = 535), while others are positive (relating to wellness) (n = 248). Coverage within SNOMED CT for negative ICNP diagnosis/outcome statements was encouraging (62%). Positive statements represented a large proportion (55%) of the complete set of unmatched concepts (n = 447).

Discussion
The value of this work extends beyond a simple assessment of content coverage. As a result of the activity 126 new concepts have been added to SNOMED CT (January 2015 release). For example, a new concept ‘deficient knowledge of symptom (finding)’ were added to SNOMED CT as an equivalent to the ICNP concept ‘lack of awareness of symptoms’. The validation activity has served also to enhance significantly the quality of both terminologies, for example through greater clarity of meaning for individual statements. For instance, the ICNP concept ‘effective verbal communication’ was renamed to ‘able to communicate verbally’ (in line with SNOMED CT syntax) as the adjective ‘effective’ in ICNP is semantically equivalent to ‘able to do’ (note that any changes in either terminology can be tracked through their respective history mechanisms). Finally, the validation activities (in addition to activities around the identification of equivalencies) have resulted in an effective way of working that will support further planned work between IHTSDO and ICN on the representation of nursing interventions within SNOMED CT. Both terminologies continue to evolve and the collaborative work is ongoing. The updated results for the most recent releases of ICNP (2015 release) and SNOMED CT (July 2015 release) will be presented at the conference.

Acknowledgements
The authors would like to thank both ICN and IHTSDO for their support, and acknowledge the contributions of Maria Braithwaite, Amy Coenen, Kay Jansen, Susan Matney, and Zac Whitewood-Moores.
References


Evaluating the Accuracy of Automated Notifiable Condition Detection in Free-Text Electronic Laboratory Report Results Using Contemporary Text Mining and Machine Learning Methods

Uzay Kirbiyik, MD, MPH\textsuperscript{1,2}, Patrick T. Lai, MPH\textsuperscript{2,3}, Brian E. Dixon, MPA, PhD\textsuperscript{1,2}, Shaun Grannis, MD, MS\textsuperscript{2,4}, Suranga N Kasthurirathne, BEng\textsuperscript{2,3}

\textsuperscript{1}IU Richard M. Fairbanks School of Public Health at IUPUI, Indianapolis, IN; \textsuperscript{2}Regenstrief Institute, Indianapolis, IN; \textsuperscript{3}IU School of Informatics and Computing, Indianapolis, IN; \textsuperscript{4}IU School of Medicine, Indianapolis, IN;

Introduction

Electronic Laboratory Reporting (ELR) refers to the process of electronically transmitting laboratory reports that identify reportable conditions from laboratories to public health stakeholders, and has been shown to improve timeliness and completeness of disease reporting.\textsuperscript{1,2,3,4} The volume of electronic reporting to state agencies is expected to increase given Stage 2 “meaningful use” program incentives from the U.S. Centers for Medicare & Medicaid Services that require eligible hospitals and encourage eligible providers to submit notifiable disease laboratory results to public health agencies using ELR.\textsuperscript{5} However, the process of identifying positive cases typically requires human review which takes time and it is infeasible both for clinical providers to manually review increasing volumes of plaintext laboratory reports and also for health departments to sift through all potentially reportable items to identify true cases. Here we demonstrate an approach for automatically identifying notifiable disease from plaintext electronic laboratory results using Salmonella as an exemplar condition.

Methods

We received 1,586 ELR’s for Salmonella through our notifiable condition detector (NCD) system from August 1\textsuperscript{st}, 2010 to July 31\textsuperscript{st}, 2012. A total of 286 unique laboratory result messages were identified after removing duplicates. We manually reviewed these laboratory messages and labeled them as either positive or negative for Salmonella. Informed by methodology developed by SNK, we removed stop words and stemmed the remaining words to identify a complete list of single-word tokens from all reports. For convenience, only tokens with an overall frequency of 3 or higher were retained and analyzed using Weka (figure 1).\textsuperscript{6} We used an information gain feature selection method\textsuperscript{7} to first rank and select tokens. We then evaluated logistic regression and random forest classifiers using the selected feature (tokens) using 10-fold cross-validation for testing the models.

Results

Text mining of 174 manually-reviewed positive and 112 negative ELRs revealed 1,102 unique tokens (out of 10,826 words). Further filtering revealed 333 tokens for our final set. After ranking by information gain, we choose the first 4, 10, 19, and 38 tokens for analysis. Our results show both random forest and logistic regression classifier methods successfully predicted the report results compared with manual review (Table 1).

Discussion

Both models exhibit higher accuracy than the current operational notifiable condition detection (NCD) system.\textsuperscript{8} Thus we conclude that it is both feasible and effective to apply text mining methods and machine learning approaches to support automated detection of public health notifiable diseases. The standard language and focused vocabulary used in the results sections may contribute to the high accuracy observed, and thus our future work will seek to evaluate the effectiveness of these methods across several other high-value notifiable diseases. These results suggest an automated system for ELR result decisions is promising; quicker and improved accuracy of automated disease detection will help not only public health and clinicians but hospital systems for their reports and planning.
Figure 1. An overview of our analysis of the notifiable disease reporting system. Lab results are transmitted via the HIE to the notifiable condition detector (NCD). Lab report transactions indicating a salmonella test were then sampled and manually labeled as positive or negative. The reports were preprocessed and discriminating tokens selected. Tokens served as features in both logistic regression and a random forest decision models, and their accuracy evaluated.

Table 1. Change on the model accuracy with the number of tokens included. Tokens were grouped based on clustering of their rank scores.

| Model accuracy | Classifier: | | | Classifier: | | |
|----------------|------------|------------|------------|------------|------------|
| Tokens used    | Logistic   | Random      | Logistic   | Random      |
|                | Sensitivity| specificity| ROC        | Sensitivity| specificity| ROC        |
| First 4        | 0.951      | 0.926      | 0.927      | 0.951      | 0.926      | 0.928      |
| First 10       | 0.962      | 0.941      | 0.962      | 0.965      | 0.946      | 0.961      |
| First 19       | 0.955      | 0.928      | 0.979      | 0.958      | 0.935      | 0.984      |
| First 38       | 0.944      | 0.913      | 0.970      | 0.962      | 0.941      | 0.979      |

References

Ill-formed Sentence Identification and Entity Extraction in Clinical Notes

Aman Kumar, Ph.D.¹, Michael J. Cairelli, DO, Ph.D.², Hassan Alam, MBA¹, Rahul Kumar, MS¹, Shweta Sheel, MBBS¹,³
¹BCL Technologies, San Jose, CA; ²Wright State University, Dayton, OH; ³Guwahati Medical College Hospital, Guwahati, Assam, India

Abstract
In this study we develop a text processing and reporting tool that processes clinical notes and identifies ill-formed constructions and extracts medical named entities. The medical named entity extraction score improves when the ill-formed constructions are identified and eliminated from the test dataset. We take cues from social media analytics studies and employ supervised machine learning algorithms and several rule based engines.

Introduction
Clinical notes have been analyzed in greater detail to harness important information for clinical research and other healthcare operations, as they depict rich, detailed medical information. For example, entity extraction systems process clinical text and extract instances of named entities that can be further grouped into specific semantic categories [1], [2], [3] and others. However, there is no significant study or analysis of processing ill-formed sentences in medical notes toward developing a system that identifies ill-formed constructions. Authors in [4] provide guidelines for parsing clinical text based on an annotated corpus to process syntactically ill-formed sentences. Other approaches such as [5] and [6] focus on increasing the robustness of the parser by enhancing accurate semantic representation. Also, there has been some work done in the field of social media on ill-formed text. Authors in [7] analyze Twitter data for text normalization using SVM and heuristic-based methods. We present an exploratory study into (1) identifying ill-formed constructions and (2) preparing a medical report based on clinical documents. The ill-formed identification module extracts text that is syntactically incorrect. The medical reporting module produces medical named entities (MNEs) that are presented chronologically in a tabular format. In this study we show that the MNE extraction scores improve if we identify ill-formed constructions in the input text. We define ill-formed constructions as constructions that are syntactically and typologically incorrect. For future work, we plan to improve the identification score and normalize the ill-formed phrases to correct phrasing.

Method
Data. We collected clinical notes - patient discharge summaries and doctor’s notes from i2b2 datasets (https://www.i2b2.org/NLP/DataSets/). We divided the datasets (2100 sentences) into randomly selected 70/30 splits for training and testing respectively. All text documents were tokenized into sentences for further processing. These sentences were manually annotated by a physician for three categories: (1) MNE-m – medical entity depicting medication; (2) MNE-s – medical entity depicting symptoms; (3) MNE – all other medical entities.

MNE Extraction. We use a combination of machine learning classifiers along with some rule-based methods. We developed a supervised linear-chain Conditional Random Fields (CRF) model based on the Stanford Named Entity Recognition (NER) system using 10-fold cross validation on the training set data. All the previously mentioned classification categories are encoded with the Begin-Inside-Outside (B-I-O) scheme. We extracted lexical and morphological features (n-gram tokens), syntactic features from the parse and dependency trees. In case of multiple MNEs, we selected MNE closest to the time entity in the parse tree. We curated a dictionary of medical phrases and abbreviations from UMLS and SNOMED CT and on-line resources (http://www.delmarlearning.com). Additional features (from [8]) specific to dosage such as mode feature set, frequency feature set etc. were also used.

Ill-formed Text Handling. In addition to the previously curated dictionary, we used aspell (http://aspell.net/) to decide outside-of-vocabulary (OOV) and inside-vocabulary (IV) terms. We extracted dependency graph features from a parse tree, in addition to N-gram features. Contextual features were key to understanding the ill-formed tokens. In addition, we modified a linear kernel SVM liblinear [9] for making decisions on ill-formedness. For ill-formed constructions, tokens which were not part of the dictionary were considered ill-formed. Also, word patterns which have periods like “p.o. q. day” were considered as IV, because they may be representing dosage etc. Similarly, words with #,@,/ like “Insulin 70/30” etc. were also considered as IV. An example of OOV constructions is: She was initially continued on Cipro/Zosyn.
**Experiment Design.** We designed three experiments using the I2B2 dataset. In the first experiment using ill-formed identification module we extracted ill-formed text from the input text. In the second experiment we extracted MNEs without extracting out the ill-formed constructions. In the third experiment we manually eliminated the ill-formed constructions first, and then processed the ‘clean’ text using the MNE extraction module.

**Results**
An illustration of MNE extraction output is given in the Table 1. For paucity of space we only show MNE extraction and not MNE-m or MNE-s extractions.

**Input Text:** September 11, 1994, coronary artery bypass surgery. He came to the emergency room with chest pain on July 22, 1994, relieved with sublingual nitroglycerin. The patient in September of 1993 had 3-4 millimeter ST segment elevation in V1 through V4 and depression in II, III, and AVF.

<table>
<thead>
<tr>
<th>MNE</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>coronary artery bypass surgery</td>
<td>September 11, 1994</td>
</tr>
<tr>
<td>chest pain</td>
<td>July 22, 1994</td>
</tr>
<tr>
<td>3-4 millimeter ST segment elevation in V1 through V4</td>
<td>September of 1993</td>
</tr>
<tr>
<td>depression in II, III, and AVF</td>
<td>September of 1993</td>
</tr>
</tbody>
</table>

Preliminary results of this study are depicted in Table 2. We use standard Recall, Precision and F-score as evaluation metrics. The first column in Table 2 depicts MNE extraction for dataset for which ill-formed-sentences are not identified and eliminated (Experiment 2). The second column gives MNE scores for the dataset for which ill-formed construction are eliminated (Experiment 3). The third column provides performance of medical entity extraction by Patrick and Li [8]. The fourth column provides ill-formed identification and extraction scores (Experiment 1). We compare this work with [8] because they record highest overall precision in the extraction of phrase level horizontal evaluations category in the I2B2 2009 medication challenge and the method is very similar to our work.

**Table 1.** MNEs with date output (Patient Report based on doctor’s notes and patient discharge summaries)

<table>
<thead>
<tr>
<th>MNE</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>coronary artery bypass surgery</td>
<td>September 11, 1994</td>
</tr>
<tr>
<td>chest pain</td>
<td>July 22, 1994</td>
</tr>
<tr>
<td>3-4 millimeter ST segment elevation in V1 through V4</td>
<td>September of 1993</td>
</tr>
<tr>
<td>depression in II, III, and AVF</td>
<td>September of 1993</td>
</tr>
</tbody>
</table>

**Table 2.** Medical Named Entity and Ill-formed text extraction results

<table>
<thead>
<tr>
<th></th>
<th>MNE Extraction with ill-formed</th>
<th>MNE extraction without ill-formed</th>
<th>Patrick &amp; Li [8] (Med entry category)</th>
<th>Ill-formed Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>82.34%</td>
<td>84.41%</td>
<td>89.62%</td>
<td>64.12%</td>
</tr>
<tr>
<td>Recall</td>
<td>71.89%</td>
<td>72.23%</td>
<td>81.38%</td>
<td>59.89%</td>
</tr>
<tr>
<td>F1</td>
<td>76.76%</td>
<td>77.84%</td>
<td>84.88%</td>
<td>61.93%</td>
</tr>
</tbody>
</table>

**Discussion**
In this study we developed a medical reporting tool that extracts medical named entities and presents them in tabular format along with respective dates. Prior to extracting MNEs, this tool identifies and extracts ill-formed constructions in the input clinical notes. We show that fixing the ill-formed construction in the input dataset will yield better medical entity extraction results. However, our MNE module failed to identify longer NP strings (longer than 3 tokens) and NPs that are rarely encountered in the training set. For ill-formed text, the system does not identify medically ill-formed constructions (abnormal dosages, wrong medication and procedures etc.). In the future we will analyze ill-formed constructions with medical inadequacies and recommend possible correct options.

**References**
Human Factors of Health Information Exchange: Barriers and Facilitators to Use of the VA’s CPRS and a Regional Health Information Exchange

Andre W. Kushniruk, MSc, PhD, FACMI1, Elizabeth M. Borycki, RN, MN, PhD1, Helen Monkman, MA1, Kenneth Boockvar, MD, MS2

1School of Health Information Science, University of Victoria, Canada; 2James J. Peters Veterans Affairs Medical Center, Bronx, New York

Introduction

Currently, worldwide there is an increasing demand for interconnection between electronic health records and regional databases and health information exchange organizations (RHIOs)1,2. The issue of how to evaluate such integration among dissimilar systems from a human factors perspective has arisen and is addressed in this podium presentation. Issues related to usability and workflow have been noted3,4,5. In addition, it has also been noted that if not done carefully, integrating systems may inadvertently lead to errors in using and interpreting information across those systems6. The objective of this study was to observe providers’ use of VA’s CPRS in conjunction with use of a RHIO tool (which were loosely integrated, requiring users to switch between the CPRS and the RHIO tool) to identify system and provider factors that impede and facilitate adoption of information exchange for routine use.

Methods

IRB approval was obtained and 12 healthcare providers were observed as they interacted with the VA’s CPRS while accessing information from the Bronx RHIO. The protocol involved audio recording the providers’ interactions with both systems in a natural environment, e.g., an office or hospital meeting room while they interacted with their own patient cases. Two observers took notes regarding what screens, system components and systems (i.e. CPRS or RHIO) were used. Participants entered data into the patient record system while “thinking aloud” or “verbalizing thoughts” in deciding on a course of action. Sessions took 30-50 minutes per participant. The resulting think-aloud protocols were audio taped, transcribed verbatim and linked in time with the record of participant-computer interactions. A coding scheme was developed to analyze the transcripts. Using multiple close readings, investigators performed initial independent coding of each transcript, generated a list of concepts and a coding scheme. Codes were then reviewed to determine if additional codes were necessary (e.g. RHIO understandability, content issues etc.). Content analysis was conducted and themes in the transcripts were identified.

Results

Analysis of the log files indicated a number of recurrent themes. In majority of the transcribed sessions, issues around participants having difficulty remembering their passwords and login to the RHIO prevented regular use of the RHIO. Other emergent themes included the perceived desire to have more data available in the RHIO (e.g. a more complete listing of out-patient medications and discharge summaries) to make the time expenditure for using it justified. The need for improved integration between the CPRS and the RHIO (i.e. not requiring users to switch between systems) was noted by a number of participants. Others indicated the need to improve the usability and organization of the information contained in the RHIO, including removing duplicate information and allowing for cut and paste features that would retain formatting. All participants indicated the need to streamline the interface and workflow to reduce time taken to access the RHIO through more seamless integration between systems when working with a heavy patient load. Another suggestion was for regular notification of improvements to the RHIO.

Discussion

Using our methodological approach (borrowing from human factors research) we found substantial barriers for VA providers in the course of using or trying to use a RHIO Exchange interface while using the CPRS. For some provider participants, these barriers outweighed the usefulness of the RHIO and they did not use the RHIO in the course of routine clinical care. In contrast, a minority of providers felt that the information in the HIE was important so they used the RHIO no matter what the workflow barriers were. These findings are being used to optimize the integration between the CPRS and the RHIO, as the method taken has been able to identify specific key barriers that can be targeted for removal in future iterations, with improved integration expected to lead to greater usability.
References

Collection and Documentation of Sexual Orientation and Gender Identity Demographic Data in the Electronic Health Record: The Patient Perspective

Brandyn D. Lau, MPH, CPH1,2,3, Laura Vail, MS2, Lisa M. Kodadek, MD1, Ryan Y. Shields, BA1, Danielle German, PhD3, Susan Peterson, MD1, Adil Haider, MD, MPH4
1Johns Hopkins School of Medicine, Baltimore MD; 2Armstrong Institute for Patient Safety and Quality, Baltimore MD; 3Johns Hopkins Bloomberg School of Public Health, Baltimore MD; 4Center for Surgery and Public Health, Brigham and Women’s Hospital, Boston, MA

Abstract
We assessed patient perspectives regarding collection and documentation of sexual orientation (SO) and gender identity (GI) demographic data in the electronic health record (EHR). Patients are more willing to report their SO and GI when all patients are routinely asked using standardized tools, and allow these data to be stored in EHRs if security and privacy concerns are addressed, and protections and training are in place to ensure appropriate information use and attention by providers.

Introduction
The Institute of Medicine and The Joint Commission have endorsed routine collection of sexual orientation (SO) and gender identity (GI) demographic data in clinical settings to better understand the health of lesbian, gay, bisexual and/or transgender (LGBT) patients.(1,2) While there is an abundance of expert opinion regarding the collection, documentation, and use of SO and GI data in the electronic health record (EHR), patient perspectives are largely unexplored. (3,4) The purpose of this study is to assess patient perspectives related to collection and documentation of patient SO and GI demographic data in the EHR and the use of these data within hospital emergency departments (EDs).

Methods
A purposive sample of participants was recruited for one-on-one in-depth interviews. Participants were recruited through community outreach, flyers, and social media. A semi-structured interview guide was developed by an expert panel of patients, community activists, and healthcare providers. Individual interviews were conducted by one of two researchers and audio data were subsequently de-identified, transcribed, and independently coded using the constant comparative method common to grounded theory. This study was funded by Patient Centered Outcomes Research Institute (PCORI) and approved by the Johns Hopkins Medicine Institutional Review Board.

Results
Fifty-three patient participants were interviewed including 9 lesbian, 12 gay, 12 bisexual, 2 queer, 14 straight and 4 participants who identified as pansexual, asexual or unsure; 16 of the 53 participants identified as transgender, bigender or genderqueer. Participants ranged in age from 22 to 67 (mean age 41.8 years) and 42% were white (non-Hispanic). Three main self-reporting tools were identified for collecting SO and GI demographic data from patients in the ED: paper form, electronic kiosk, and cellphone app. Patient perspectives regarding reporting of SO and GI demographics were categorized into six major themes: privacy, medical relevance, normalization, recognition, patient-provider relationship, and ED flow (Table 1). Participants were more willing to provide SO and GI demographic data when it is clear that all patients are routinely being asked in a standardized manner and when offered a rationale about why the information is needed and how the information will be used.

Recognition that SO and GI demographic data would ultimately be stored in the EHR made private reporting of this information on an electronic kiosk or smartphone app more appealing than verbal reporting during the clinical encounter. Patients commonly cited privacy concerns about storage of SO and GI demographic data within EHR systems, including access by third parties such as employers and insurance companies, as well as unintended disclosure to family and friends. Patients noted that documentation of SO and GI demographic data within the EHR may facilitate proper recognition of preferred name, pronouns, and relationships by providers throughout the ED encounter. Further, patients felt that self-reporting prevented the possibility of judgment or discrimination by the provider and alleviated the need to pre-judge the provider for cues to safety or acceptance of SO and GI minorities. Patients reported multiple benefits associated with electronic collection and documentation of SO and GI
demographics including the ability to update/modify their responses, use their own devices (e.g. cellphone app, patient portal), and having more control over own data. Commonly, patients cited privacy concerns with paper-based documentation of SO and GI demographic data due to less perceived monitoring than electronic methods.

Table 1. Thematic Intersections with Collection Mechanisms: Facilitators and Barriers

<table>
<thead>
<tr>
<th>MAJOR THEMES</th>
<th>(+) = Facilitator</th>
<th>(-) = Barrier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+) More private than answering verbally</td>
<td>(+) Demonstrates that hospital/clinic cares about SO and GI</td>
<td></td>
</tr>
<tr>
<td>(+) Silent transmission of data</td>
<td>(+) Facilitates societal recognition of SO and GI minorities</td>
<td></td>
</tr>
<tr>
<td>(-) Others may see responses</td>
<td>(+) Alleviates need to judge provider and search for cues to safety</td>
<td></td>
</tr>
<tr>
<td>(+) Concerns about storage and use</td>
<td>(+) Possibility of having registrars/providers help fill out intake forms</td>
<td></td>
</tr>
<tr>
<td>(+) Fill out on own terms and location</td>
<td>(+) Collection as demographic information</td>
<td></td>
</tr>
<tr>
<td>(+) Revising forms to be more inclusive</td>
<td>(+) Inclusion in most tech-advanced means</td>
<td></td>
</tr>
<tr>
<td>(+) Collection as demographic information</td>
<td>(+) Collection as demographic information</td>
<td></td>
</tr>
<tr>
<td>(+) Collection as demographic information</td>
<td>(+) Inclusion in most tech-advanced means</td>
<td></td>
</tr>
<tr>
<td>(+) Collection as demographic information</td>
<td>(+) Patients self-disclose on intake</td>
<td></td>
</tr>
<tr>
<td>(+) Collection as demographic information</td>
<td>(+) Collection as demographic information</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

Research surrounding the health status of LGBT individuals in the United States is limited because SO and GI data are not routinely collected nor documented in clinical settings. National organizations advocate for collecting these data from patients, but have offered little patient-centered guidance for the most appropriate methods for collecting, documenting, and using these data. This is the first study to ascertain patient perspectives regarding the collection and documentation of SO and GI demographic data in the ED. Our study suggests that patients are more willing to report their SO and GI demographics when this information is asked routinely as part of standard policy. Furthermore, patients are generally willing to allow their SO and GI demographics to be stored in EHR systems if security and privacy concerns are addressed and protections and training are in place to ensure appropriate information use and attention by providers. Furthermore, patients consider that asking these questions during intake signals that the ED recognizes and respects sexual and gender minorities. Future work will engage a national sample of patients to develop appropriate, patient-centered methods to collect SO and GI demographic data in EDs.

References

1. Institute of Medicine (US) Committee on Lesbian, Gay, Bisexual, and Transgender Health Issues and Research Gaps and Opportunities. 2011.
Rethinking Document Retrieval for Scientific Literature: A Learning to Rank Approach
Jesse M. Lingeman¹, Hong Yu, PhD²
¹University of Massachusetts Amherst, Amherst, MA, ²University of Massachusetts Medical School, Worcester, MA

Introduction

With the rise of natural language queries from smart phones, full document queries for article recommendation, prior art for patents, or document duplication, search needs have been moving towards longer queries. Search using a full document as a query makes it helps the user as they don’t have to guess which keywords to use or form complicated Boolean queries. However, previous work has shown that simply using an entire document as a query does not work well¹. A long document is too noisy and it is difficult to identify the important parts of a document. Previous work has used translation models²,³, collaborative filtering⁴, NLP techniques⁵, topic models⁶, and other approaches.

In this work we approach query document search from a learning to rank perspective. Learning to rank works as follows: An initial query is run to retrieve a subset of the corpus, and then a feature vector is generated between the query and each document in the subset. The feature vector is a list of measures relating the document to the query. Learning to rank algorithms then build a supervised model that ranks a set of judged documents.

The primary contributions of this study are as follows: 1) Introduce a novel recommendation task with its own unique challenges. 2) Provide a simple yet effective framework for crowdsourcing tasks that require a domain expert to perform. 3) Apply a learning-to-rank framework to the task of query document search.

Methods

We apply our model to a crowd-sourced dataset. Given the full text of an academic publication, we want to rank its citations in order of importance. We collected recent publications from PLOS journals and asked the authors to rank by importance 5 citations we selected from their recently published paper. In total we collected 90 fully annotated publications. For each document and its citations we calculate text similarity features, citation position features, and the age of the citation. Citations were ranked using the LambdaMART⁷ implementation in the RankLib package from the Lemur toolkit⁸ and with SVMRank⁹. Normalized discounted cumulative gain (NDCG) is used as our performance metric. NDCG penalizes incorrect ranking of highly relevant documents more harshly than less relevant document.

Results

We first introduce two baselines: 1) a random baseline where we randomly shuffle the relevance judgments and 2) a model that uses the original pseudo-relevance judgments we used to select the citations to send to the authors. We find from testing that our random baseline has an average NDCG of 0.73. The data are split into 100 random train/test splits with 80% training data and 20% test data. SVMRank obtains an NDCG of 0.808 using the annotated data and 0.792 using the pseudo-relevance judgments (p < 0.05). LambdaMART obtains an NDCG of 0.81 using the annotated data and 0.75 from the pseudo-relevance judgments (p < 0.05). Feature selection is performed; with the best performing feature sets give SVMRank an NDCG of 0.824 using the annotated data and 0.808 using pseudo-relevance judgments. LambdaMART’s NDCG does not increase after feature selection.

Discussion

We have found that the learning to rank approach is effective for treating documents as queries. The model trained using the author’s judgments does significantly better than the model trained using the pseudo-relevance judgments.

We found that the most important features for both models are how frequently a citation is referenced in an article, the age of the citation, and the text similarity of the citation’s abstract with the discussion/conclusions sections of the article. The newer a citation is the more likely it is to be directly related to a the article rather than an older, more foundational citation, and restricting text similarity to discussion/ conclusion helps reduce noise in the document.

What we have shown is that the learning to rank approach, while requiring some feature engineering, can be applied to the task of query document search. In future work we will expand this approach to larger datasets across varied query document search datasets.
References

Identifying Novel Adverse Drug Events from Health Social Media Using Distant Supervision

Xiao Liu, PhD student¹, Hsinchun Chen, PhD¹
¹University of Arizona, Tucson, Arizona, United States

Introduction

Adverse drug events (ADEs) are a significant health problem worldwide. Many prior studies have extracted ADEs with supervised learning and unsupervised learning approaches. Distant supervision leverages existing knowledge and requires no manually labeled text for learning relations, which has not been explored for ADE extraction before. We develop a framework for identifying novel adverse drug events from social media using distant supervision. We utilize FDA’s Averse Drug Event Report System (FAERS) as a knowledge base for distant supervision and predict relations between medication and medical conditions in social media text.

Methods

Figure 1 shows the research method we propose for identifying novel adverse drug events from health social media using distant supervision. The data for this study comes from three sources: health social media discussion forums, Twitter, and the FDA’s adverse drug event reports from FAERS. We choose diabetes treatments as a research case to test the efficacy of our framework. We conduct data preprocessing with the following steps: medical entity extraction, co-reference resolution, and syntactic parsing. Medical entity extraction mainly aims to identify medical condition mentions and drug mentions health social media. Syntactic parsing generates features for distant supervision. We incorporate co-reference resolution to link medical entities across sentences within the same document. We develop our distant supervised learning for adverse drug event detection model based on the multi-instance multi-label assumption¹. We assume that each relation mention involving an entity pair has exactly one label, but allows pair to exhibit multiple labels across different mentions. It can be defined as a function that takes health social media data collection \(C\), a set of entity medical entity mentions extracted from \(C\) \(e\), and an extraction model, and outputs a set of relations \(R\) such that any relation extracted has at least one positive instance in \(C\). To train the model, we use FAERS as a knowledge base for drug indication and adverse drug event relations \(D\). Using distant supervision, \(D\) is aligned with sentences in \(C\), producing relation mentions for all relations in \(D\). EM algorithm is used to train the classifier. We evaluate our framework by held-out evaluation and manual evaluation. Held-out evaluation is conducted automatically by holding out part of the FAERS relation knowledge during training and comparing newly discovered relation instances against this held out data. We choose the single-instance single-label (SISL) model², and the multi-instance single-label model (MISL)³ as baselines for this automatic evaluation. This evaluation focuses on accuracy relation labels of each entity pair. For human evaluation, we evaluate the accuracy of relation labels on mention level. Among the 41 diabetes treatments, three of them were approved in the past two years: Afrezza in June 2014, Farxiga in January 2014, and Invokana in January 2013. We manually examine their top ranked relation labels.

Results

Table 1 shows a summary of the data after preprocessing. Overall, we retrieve about 17,000 sentences with at least one entity pair. These sentences contain 37 unique medication entities and 476 unique medical conditions.

Table 1. Summary of medication and medical condition mentions in data

<table>
<thead>
<tr>
<th>Medication</th>
<th>Medical Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>476</td>
</tr>
</tbody>
</table>
In Figure 2, we can see that MIML model has a larger area under precision-recall curve than MISL model and SISL model which indicates MIML model can predict relation labels with higher performance. This evaluation may under estimate the accuracy. If a true positive adverse drug event doesn’t exist in knowledge base, it will be considered as “no relation” in evaluation. Besides, some medical conditions and medications may have been reported to FDA but their mentions in health social media more often present “no relation”.

Figure 2. Precision recall curve for distant supervision

Table 2 shows the top ranked relations we found for Afrezza, Farxiga, and Invokana. Most of them are adverse drug events. These adverse events of medications new to the market can contribute to understand and assess the risk of new medications for both patients and clinical practitioners.

Table 2. Top ranked relation predictions for Afrezza, Farxiga, and Invokana

<table>
<thead>
<tr>
<th>Mediation</th>
<th>&lt;Relation label&gt;: medical condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrezza</td>
<td>&lt;cause&gt; oral bleeding; &lt;cause&gt; hypoglycemia; &lt;cause&gt; cough; &lt;cause&gt; lung problem; &lt;cause&gt; overdose; &lt;cause&gt; irritation; &lt;cause&gt; pain</td>
</tr>
<tr>
<td>Farxiga</td>
<td>&lt;cause&gt; weight loss; &lt;treat&gt; diabetes; &lt;cause&gt; bladder cancer; &lt;cause&gt; breast cancer; &lt;cause&gt; hypotension; &lt;cause&gt; genital yeast infections; &lt;cause&gt; tiredness; &lt;cause&gt; running nose</td>
</tr>
<tr>
<td>Invokana</td>
<td>&lt;cause&gt; bladder cancer; &lt;cause&gt; hypotension; &lt;treat&gt; diabetes; &lt;cause&gt; kidney disease; &lt;treat&gt; urinary tract infections; &lt;cause&gt; weakness; &lt;cause&gt; hypoglycemia; &lt;cause&gt; weight gain</td>
</tr>
</tbody>
</table>

Discussions

In this research, we develop a research framework to identify adverse drug events from health social media using distant supervision. This is the first study that incorporates distant supervision in health social media ADE detection. By aligning knowledge base data with health social media data, we are able to generate a large number of data for training at low cost and the performance is promising. It can also help to evaluate the similarity and difference between health social media reports and FAERS reports. In the future, we plan to expand our research test bed to include more drugs. We will evaluate our approach with comparison to rule-based approach, and supervised learning approach for extracting adverse drug events.

References

Design Principles for Clinical Decision Support for Direct Use by Patients: Addressing Symptom Self-Management in Cancer Patients

David F. Lobach, MD, PhD, MS1; Janet L. Abraham, MD2; Donna L. Berry, RN, PhD2; Michael S. Rabin, MD2; Ilana Braun, MD2; Manan Nayak, MA2; Mary E. Cooley, RN, PhD2

1Klesis Healthcare and Duke University Medical Center, Durham, NC; and 2Dana-Farber Cancer Institute, Boston, MA

Introduction
Direct use of clinical decision support (CDS) by patients, particularly in the area of symptom management, has not been studied extensively. Only four studies described rule-based CDS tools for direct use by cancer patients.1-4 None of these studies has specifically focused on the patient barriers and clinician concerns that need to be addressed to create safe, effective CDS for patients. The purpose of this study was to identify information patients need and their preferences about being informed in order to create CDS that can empower and engage patients to self-manage their cancer-related symptoms. This study extends the literature by identifying design principles for creating CDS tools for managing cancer symptoms. We postulate that some of the principles we identified may be useful for developing CDS for patients across diverse clinical scenarios.

Methods
Our multi-disciplinary research team conducted a formative mixed methods study involving focus groups, semi-structured interviews, and usability surveys (Acceptability E-scale)5 with cancer patients, their caregivers and oncology clinicians. All focus group and interview discussions were audio-recorded and content was transcribed for analysis. We elected to focus on the most common symptoms contributing to emergency and urgent care visits among patients with cancer, namely pain, constipation, and nausea and vomiting.6 Based on the data gathered and on input from stakeholders participating on expert panels that included patients, caregivers, clinicians, information system developers, and administrators, we iteratively created prototypes suitable for display on desktop computers or mobile devices to simulate design and content features of a self-management CDS system for patient symptoms to which study subjects could respond. The conceptual framework of this project was informed by integration of the Technology Acceptance Model7, addressing system function and encompassing the perceived usefulness and perceived ease-of-use factors, with the refined version of the Chronic Care Model8,9, including elements of patient safety, cultural competency, care coordination, and community policies regarding community resources.

Results
For the formative evaluation, we recruited 24 cancer patients who had experienced the symptoms of interest within the past 6 months and 13 oncology clinicians from the Dana-Farber Cancer Institute. Through this design and evaluation process, we identified patient barriers and clinician concerns to using CDS for symptom self-management. To address these barriers and concerns, we formulated CDS design principles for patient safety, cultural competency, care coordination, resource availability, and system function (Table1). Patient usability surveys indicated strong support for the CDS tools.

Discussion
The findings from this study provide insight into how CDS systems can be developed to support patient self-management directly. Patient safety and tool navigation were critical features. Many of the insights gleaned from this study may be generalizable and could inform the development of CDS resources that patients with other diseases and their caregivers can use for self-management of symptoms. The barriers and concerns we identified for CDS for symptom management in cancer patients had both similarities and differences to barriers associated with self-management of non-malignant chronic conditions.10,11 Similar barriers included lack of knowledge, poor communication between patients and providers, and logistical issues for obtaining care. Other differences found in our study of CDS for cancer symptom self-management but not found in patients with non-malignant chronic conditions were patient safety, including the need to detect life-threatening problems and uncertainty about the safety of clinical recommendations; health beliefs that attribute negative connotations to certain recommendations; and issues related to CDS system function and navigation.

Conclusion
In this project, we identified design principles to enable creation of an evidence-based CDS that both engages cancer patients to self-manage symptoms using pharmacologic and non-pharmacologic therapies and clearly indicates when the patient should seek the help of a clinician.

Acknowledgements
This project was funded in part by The Patient Centered Outcomes Research Institute Grant PI-12-001.
<table>
<thead>
<tr>
<th>Barrier or Concern</th>
<th>Issue</th>
<th>Design Principle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Safety</td>
<td>Identification of Life-Threatening Problems</td>
<td>Err on the side of caution when patient safety is in question</td>
</tr>
<tr>
<td></td>
<td>Provide explicit instructions for patients regarding contacting clinicians about concerns</td>
<td></td>
</tr>
<tr>
<td>Uncertainty about Following Recommendations</td>
<td>Inquire about appropriateness of recommendations prior to offering advice</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Encourage patients to communicate with clinical team about interventions</td>
<td></td>
</tr>
<tr>
<td>Consensus on Best Practices</td>
<td>Build algorithm content based on established guidelines and best practices</td>
<td></td>
</tr>
<tr>
<td>Cultural Competency</td>
<td>Unfamiliar Clinical Terminology</td>
<td>Test word selection with intended end-users</td>
</tr>
<tr>
<td></td>
<td>Make questions explicit and detailed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Enhance communication with graphics, especially for clinical concepts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provide lists to enable user to identify specific items such as medications</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Designate consistent presentation areas for consistent display of specified type of information</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Use the most simple, unambiguous language possible to convey concepts</td>
<td></td>
</tr>
<tr>
<td>Lack of Clinical Framework</td>
<td>Provide educational information to promote understanding</td>
<td></td>
</tr>
<tr>
<td>Health beliefs</td>
<td>Identify health beliefs that may impact interpretation of content and modify content accordingly</td>
<td></td>
</tr>
<tr>
<td>Care Coordination</td>
<td>Patient-Clinician Communication</td>
<td>Encourage patients to notify their clinical care team about interventions that they have followed</td>
</tr>
<tr>
<td></td>
<td>Patient activation</td>
<td>Provide explicit, detailed, actionable instructions to the extent possible</td>
</tr>
<tr>
<td></td>
<td>Determine what users are willing to do prior to making recommendations</td>
<td></td>
</tr>
<tr>
<td>Community Policies/Resources</td>
<td>Clinical Resources</td>
<td>Determine what resources are available to the patient</td>
</tr>
<tr>
<td></td>
<td>Technology Resources</td>
<td>Create tools that will function across multiple platforms</td>
</tr>
<tr>
<td>Technology Acceptance</td>
<td>Data Entry</td>
<td>Streamline data entry</td>
</tr>
<tr>
<td></td>
<td>Provide comprehensive set of selection options</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Navigation</td>
<td>Optimize workflow</td>
</tr>
<tr>
<td></td>
<td>Track progress for user</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accommodate user changes and pauses</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provide context for all interactions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ensure completeness and uniqueness of pathways through algorithm</td>
<td></td>
</tr>
<tr>
<td>Patient Engagement</td>
<td>Personalize content</td>
<td></td>
</tr>
<tr>
<td>Accessibility</td>
<td>Enhance readability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accommodate multiple platforms</td>
<td></td>
</tr>
</tbody>
</table>

References

Introduction

The surveillance of sexually transmitted diseases (STDs) is an important disease control activity of the Centers for Disease Control and Prevention (CDC). STDs contribute a significant burden of disease morbidity for Americans, even though they are largely preventable and treatable. Left undiagnosed and untreated, the costs can be high and include intractable pelvic inflammatory disease, infertility and infant death. Fifty-nine state and territorial jurisdictions report STDs that are on the list of national notifiable diseases to the CDC’s Division of STD Prevention (DSTDP). These diseases were responsible for over 1.8 million cases of infection reported to the CDC in 2013 (CDC, 2014).

STD Programs at the state and local levels report using electronic information systems for surveillance that meet the CDC’s NEDSS (National Electronic Disease Surveillance System) standards. For the past 25 years, DSTDP has supported the application STD*MIS for the states to use for this purpose. STD*MIS also provides functions for case management and partner services (PS). However, DSTDP is no longer supporting this system and the trend is towards interoperability across infectious disease reporting. Thus, states are migrating to integrated systems.

As of December 2014, 22 states and territories continued using STD*MIS to collect and transmit STD data for surveillance and PS; several others use it only for PS. These jurisdictions are in the process of selecting new systems, and many have requested advice and technical assistance from DSTDP. Of the jurisdictions not using STD*MIS, 10 used PRISM®, 7 MAVEN®, 3 TriSano® and 17 had custom made systems.

Methods

The study was designed to gather lessons learned about selection and implementation from states that have already transitioned to one of three off-the-shelf (OTS) systems, MAVEN®, PRISM® and TriSano®. Semi-structured interviews were conducted with STD program administrators for seven jurisdictions about their experiences selecting, contracting for and implementing a new information system for STD surveillance and program management. Selection of jurisdictions to be interviewed was based on recommendations from DSTDP staff who work with the local jurisdictions using the three systems of interest. The primary author had no previous experience with the interviewees or the systems under study.

Results

Phone interviews were conducted with 3 jurisdictions using MAVEN®, 2 using PRISM® and 2 using TriSano®. All of the jurisdictions had begun their transition to these systems in the last seven years with the most recent being in 2013. Respondents had been using custom database systems or STD*MIS prior to adopting the new system. The interviews covered selection and deployment, components and satisfaction, impact, and general lessons learned. Interviews were recorded in hand-written notes augmented by review of audio recordings. A case study was developed for each interview. The case studies were analyzed qualitatively to identify lessons learned. While two of the interviews were with one person involved in administering the STD Program, most of the interviews were with a group of people who were involved with the system either in the STD Program or as Information Technology (IT) support. General lessons learned across systems and implementations are summarized in Table 1 below. The data suggest that the differences between systems are more due to the implementation process than OTS. However, further study is needed to directly compare the functions and usability of each software package.

Conclusion

The experiences of public health department jurisdictions that have transitioned to new information systems in recent years, can inform others who are currently in the planning process. While individual systems may have their strengths and deficiencies, the process of adopting a new system can be managed in a way that improves the process of implementation and the impact of the deployment on data quality and program activities. Our results suggest that
jurisdictions take their time not only to plan for the implementation, but to develop relationships with IT departments and vendors, consider funding streams for maintenance of the system and prepare staff for new roles and responsibilities.

Table 1. General Lessons Learned from Interviews with STD Program Administrators from 7 jurisdictions

<table>
<thead>
<tr>
<th>Selection/Deployment</th>
<th>Components/Satisfaction</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talk with jurisdictions that have already been through this process.</td>
<td>All of the systems adequately meet basic surveillance and PS requirements.</td>
<td>Data Quality: Generally, timeliness of data is improved. However, with ELR, data volume increases; implementations vary in automatic de-duplication abilities.</td>
</tr>
<tr>
<td>Involve STD Program staff in selection of the system and contracting including specifics in the scope of work for functions critical to STD that may not be for other public health programs.</td>
<td>Areas for improvement:</td>
<td>Work Flow: Electronic data receiving may mean data entry staff must learn new skills. Changes in job duties could mean new positions and higher salaries are needed. In some cases, duties were shifted to local health departments from the state.</td>
</tr>
<tr>
<td>Expect the deployment to take time. Plan the process. Test the system thoroughly with users before it goes live. Involve people at multiple levels in the planning and deployment process.</td>
<td>• flexibility in reports; • analytics and mapping; • automatic linking and pre-populating for new cases when a person is already in the system; • ELR integration</td>
<td>Funding: New systems require close relationships with IT which has implications for how these services are funded across programs.</td>
</tr>
<tr>
<td>Foster a good working relationship with IT personnel.</td>
<td>System usability and user-friendliness is as much a function of internal IT support and vendor maintenance as the systems themselves.</td>
<td></td>
</tr>
</tbody>
</table>

References

Leveraging Business Rules Techniques for Data Quality Assurance in Public Health Systems

David Lyalin, PhD and Warren Williams, MPH
National Center for Immunization and Respiratory Diseases (NCIRD), Centers for Disease Control and Prevention (CDC), Atlanta, GA, USA

Abstract
This presentation discusses a case study of implementing business rules for validating quality of data coming to immunization information systems that operate within U.S. state immunization programs. Independent evaluation findings indicate that application of developed best practice recommendations in state and local health departments resulted in improved data quality, reduced staff time, and increased efficiencies across immunization programs.

Introduction
This presentation discusses a case study of implementing business rules for validating quality of data coming to immunization information systems (IIS) that operate within U.S. state immunization programs. The accuracy, completeness, and timeliness of immunization records in IIS is a decisive factor in improving the health of patients, the operation of healthcare clinics, and public health decision making.

Methods
A panel of immunization experts has been assembled to develop consistent business practices for validating various streams of immunization and demographic data coming to IIS. Business rule technique in combination with concept modeling and other analysis techniques were successfully used during in-person facilitated sessions and teleconferences to discuss, formulate, and document consensus-based recommendations. Special attention has been paid to alignment between business strategies (principles – high-level direction) and business rules, as well as to actions on business rules violation and classification (grouping) scheme for business rules that follows combinations of key elements of the concept model, such as Patient Date of Birth, Vaccination Encounter Date, and Vaccine Type.

Results
Business rules helped to analyze immunization tracking operations and document consensus-based best practice recommendations for advancing quality of immunization data. The resulting best practices guide helps immunization information systems identify common errors in the incoming data streams using standardized business rules for validation checks and decision-making. Independent evaluation findings indicate that application of these best practice recommendations in the IIS domain of state and local health departments resulted in improved data quality, reduced staff time, and increased efficiencies across immunization programs. Most immunization programs (78%) were familiar with these recommendations; 76% of IIS used best practice guidelines. Users of the guide reported clear advantages to not “reinventing the wheel” – business rules that reflect best practices could be directly adopted or modified as needed, which resulted in reduced IIS staff time.

Discussion
Our experience indicates that collaboration among stakeholders and systematic analysis with business rules techniques promote system thinking and improves operations of the immunization registration area of public health. It also helps break otherwise complex programmatic and operational challenges into manageable fragments which can be incrementally improved upon.

References
Early Experiences with Meaningful Use and Online Portal Implementation among Providers/Staff and Patients/Caregivers in a Safety Net Healthcare System

Courtney Lyles, PhD, Lina Tieu, MPH, Dean Schillinger, MD, Neda Ratanawongsa, MD, MPH, Urmimala Sarkar, MD, MPH

University of California, San Francisco, Division of General Internal Medicine at San Francisco General Hospital, Center for Vulnerable Populations, San Francisco, CA

Introduction:
Online portals, which provide patients access to electronic health record information, are becoming ubiquitous in the US, especially with the support of Meaningful Use financial incentives. Because of their widespread implementation across many healthcare systems, portals will be the first health technology to reach many diverse patient populations across the country, and they will continue to be a platform through which we integrate future mHealth solutions (e.g., data from mobile phone apps and sensors) in the coming years. However, there is very little information about the reality of implementing portals in safety net systems that face higher numbers of competing demands and more limited resources. While our previous work has found high interest in electronic communication with providers among a diverse group of low-income patients in our healthcare setting, we sought to use qualitative methods to elucidate real-world experiences with portal implementation in the few months directly preceding portal implementation.

We conducted formative assessments with both providers/staff and patients/caregivers.

Methods:
This study took place within the San Francisco Health Network (SFHN), which includes 11 community-based primary care clinics as well as 4 hospital-based primary care clinics at San Francisco General Hospital. In this setting, the majority of patients served are low-income and publicly insured.

In 2014-2015, we conducted two rounds of data collection with both providers/staff and with patients: 1) a survey of 77 primary care providers and staff (January-February 2015), and 2) 16 in-depth interviews with chronic disease patients or their caregivers (from December 2013 to August 2014). The provider/staff survey was sent to all providers (physicians, nurse practitioners, nurses, medical assistants) and staff (front desk staff, etc.) in primary care across SFHN. This online questionnaire included closed-ended items about recommending portal use to patients as well as required open-ended items about expected barriers in using the system, particularly in using secure messaging through the portal. Patient interviews were focused on a more narrowly defined group of individuals who were most likely to use the portal website when it was launched in 2015: English speakers who expressed at least some interest in using the Internet for healthcare management (as the portal is not yet available in other languages besides English). These one-on-one in-depth interviews explored topics of experiences with the healthcare system, technology use in everyday life, and interest in using an online portal to manage healthcare tasks. We showed each participant example screenshots of the portal website during the interview.

We used both descriptive statistics and inductive, open coding to assess barriers and facilitators to portal use.

Results:
The provider/staff survey sample included 46 physicians, 18 other providers (nurse, medical assistant, etc.), and 11 staff members, with about half of respondents (45%) from community-based clinics. The mean age was 42 years, and 77% were female and 46% were white. Overall, while most stated they would recommend portal use to their patients, it was often to only a subset of their patient panel. In terms of challenges for their patients in using a portal, English proficiency (81%), literacy (75%), unreliable access to computers/Internet (71%), and unstable housing (61%) were the biggest reported perceived barriers for patients. When considering using secure messaging through the portal in their own practice, lack of time (68%) was the biggest barrier, followed by worry about patients overusing or sending irrelevant emails (47%). Specific example quotes from providers can be found in Table 1, expressing their range of feelings about incorporating portal use into their workflows in a resource-limited setting.

A total of 45 patients and caregivers were approached for individual interviews. Of those approached, 25 (59%) expressed interest in the study, 7 (16%) declined citing lack to interest in computer use, and 13 declined due to unknown or other reasons (too busy, uninterested in research). Of those who expressed initial interest, we enrolled 16 participants (55%) in the study. The patient sample included 11 patients and 5 caregivers who were existing computer users. The mean age was 55 years, 50% were African American, 63% were male, and 70% reported using the Internet daily – although the range of computer abilities/tasks was broad, with a third of participants not currently using email. (We feel that this range of computer ability matches the skill set of the broader SFHN patient population who have at least some computer experience.) The major barriers to portal use were: concerns about security of information online, lack of technical skills/interest, and preferences for in-person visits/communication. Several participants discussed fundamental barriers to using an online portal, including: challenges with reading, writing, and typing (specific quotes can be found in Table 2).

Discussion:
Providers/staff and patients/caregivers in a safety net system reported major barriers to using an online portal. Overall, there was a high level of concern about reading and understanding medical information and the security of the information online. Patients were specifically concerned about preserving existing in-person relationships with providers, and providers expressed concerns about having enough time to be able to connect with patients through secure messaging. Our findings suggest 1) a strong need for training and support to assist safety net patients with portal registration and use, perhaps most so for those with limited health literacy, and 2) considerations for changes in reimbursement or practice structures to support clinician time to be able to communicate effectively through electronic media with vulnerable patient populations.
Table 1. Provider/Staff Barriers to Portal Use

<table>
<thead>
<tr>
<th>Recommending the portal to patients</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Will recommend portal to most patients</td>
<td>I think it is important for patients to be informed of their care.</td>
<td>[The portal] will increase communication and empower patients to engage in their health care.</td>
</tr>
<tr>
<td></td>
<td>Regardless of a patient's having a computer or knowing how to use one, I think all patients have the right to know the portal exists and to try it out (if it works well...). I use Kaiser's and I find it very useful.</td>
<td></td>
</tr>
<tr>
<td>Will recommend portal to some patients only</td>
<td>I will recommend this for patients who have computer access and who will benefit from access (as opposed to getting confused and anxious).</td>
<td>For many patients this recommendation will be seen as a burden and something they are unlikely to be able to accomplish but will want to attempt in order to please the medical provider.</td>
</tr>
<tr>
<td>Unsure of whether will recommend portal to patients</td>
<td>I already have phones, jelly-beans [EHR tasks], email and mailbox to cover in answering patient needs and one more thing will drive me from primary care.</td>
<td>[I'm] not sure if it will be well used/accessed.</td>
</tr>
<tr>
<td></td>
<td>So few of our patients have access to computers and/or are comfortable working with them.</td>
<td></td>
</tr>
<tr>
<td>Won't recommend portal to patients</td>
<td>Without being able to securely contact providers to discuss results of health information this feature may not be very helpful to helping patients understand their health...I would wait until the communication piece is available before I would initiate the program.</td>
<td>For patients who do not know how to read or write in any language, such a portal will not be useful at all, further marginalizing their care.</td>
</tr>
<tr>
<td></td>
<td>[The EHR] is still full of bugs. I have little faith that this Portal will function effectively as it is housed in a poorly designed EMR system, and will only add to the confusion and tangles that our patients are confronted with.</td>
<td></td>
</tr>
</tbody>
</table>

Role of family members or caregivers in portal use

| Proxy access will be useful | I think that the adult children of some of my patients might find this useful in communicating and reaching me in a more timely manner than calling the clinic or my voicemail. I wonder how much this will help as I’ve offered my e-mail to some family members and they have declined because a majority of my patients do not know how to access e-mails or the internet. | |
| | Many of our pts would likely require the assistance of family/caregivers to use the portal due to low literacy and limited English proficiency. The extent to which this would be successful will likely depend on our ability to do outreach and education for these family members and caregivers. | |
| Concerns about proxy access | [I] worry about HIPAA. | Most [of my patients] do not have any family nor caregivers. [They are] homeless outdoors. |
| | For some this would be helpful. For others it would undermine their sense of privacy and independence. | |

Table 2. Patient Barriers to Portal Use

| Limited computer/Internet access | I go to the library sometimes or a friend’s house or something there, or when I get with the tutor or something and they're teaching me something, they'll teach me on their computer or stuff like that. | Plus, I don’t have $35.00 that happens to be the monthly fee [data for Internet on mobile phone]. |
| Fear & Experience with viruses | I’m scared of that because of viruses and stuff like that. I don’t want to. I’m a poor guy. I can’t get the computer fixed if I break it or if something infects it. | |
| Limited proficiency | I can’t type. Everything is like...I’ve got to do the two-finger thing. | That’s another thing because you got to have so many words and letter. You know, characters, so how do you distinguish that? I mean you say characters, are they letters? |
| Concerns about security and privacy | Regardless of what a person says that this site is secured and all that, I just don’t believe it....It’s not only hospitals but pharmaceutical and every researcher will tap into my information. That’s the thing that I worry about. | Hackers getting [site] everything....I had to change banks because...they had everything - my name and address - my mom's maiden name. |

References
Using PopMedNet to Support a Multi-Site Research Network: Lessons from PCORnet

Jessica M. Malenfant, MPH1, Jessica L. Sturtevant, MS1, Laura G. Qualls, MS2, Kyle Erickson, BS1, Lesley H. Curtis, PhD2, Jeffrey S. Brown, PhD1

1Harvard Pilgrim Health Care Institute, Boston, MA; 2Duke Clinical Research Institute, Durham, NC

Introduction: The National Patient-Centered Clinical Research Network (PCORnet) distributed research network (DRN) was established by The Patient-Centered Outcomes Research Institute (PCORI) in 2013. Phase 1 of the project included an ambitious 18 month period to implement a distributed network of networks supported by the PopMedNet™ (PMN) informatics platform. PopMedNet supports a range of health data networks, and with PCORnet, there are currently over 100 diverse sites utilizing the system across the various networks. The PCORnet implementation was the largest and most complex for any of the PopMedNet-based networks.

Methods: Implementing the PCORnet DRN involved rapidly establishing a secure PopMedNet system for PCORnet. PCORnet - designed as a “network of networks,” coupled with the large number of data partner nodes, differences in local network architecture, workflows, governance, diversity of partner informatics and software expertise, and the lack of experience with PMN and distributed analysis created numerous implementation challenges. Successful implementation was enabled through re-use of PMN components developed by previous networks, including components that had to be re-configured, enabled, or enhanced to ensure that the PCORnet could take advantage of the tools.

Based on work supported by the FDA Mini-Sentinel project, the existing PMN architecture was easily configured to meet the PCORnet usability, workflow, and governance requirements. Specifically, the methods used to associate related organizations with each other in the system coupled with integrated workflow processes related to the request and response tasks were leveraged for use in PCORnet.

Additionally, the PCORnet implementation made use of, and extended the PMN collaborative toolkit that includes collection and discovery of information about organizations and data resources. The collaborative tools used and extended by PCORnet were initially developed by the NIH Health Care Systems Research Collaboratory and FDA Mini-Sentinel project to improve the capture and use of institutional, data source, and request metadata.

Results: The PopMedNet instance for PCORnet was successfully implemented for use across the network, with over 70 unique installations representing a variety of network architecture configurations, governance, and workflow processes. Several reporting, analytic, and workflow tools supported and used by other distributed networks were also configured for use in the PCORnet. During this implementation and testing phase over 250 queries were distributes across PCORnet.

Discussion: The size and complexity of the PCORnet implementation created substantial challenges. However, the PMN infrastructure established and used by other distributed health data networks has proven to be scalable and extensible and allowed for rapid deployment for PCORnet. As PCORnet distributed research and querying activities increase and the network grows, there is an increasing need to coordinate and leverage work and resources across stakeholders and other PopMedNet networks.
Qualitative Study of an Electronic Tool for Facilitating Problem-Solving and Sensemaking in Diabetes Self-Management, Mobile Diabetes Detective

Lena Mamykina1, Ph.D., Elizabeth M. Heitkemper2, MSN, RN, Arlene M. Smaldone2, Ph.D., CPNP, CDE, Rita Kukafka1, Ph.D., Patricia G. Davidson3, DCN, RDN, Elizabeth D. Mynatt4, Ph.D., Jonathan N. Tobin5, Ph.D., Andrea Cassells5, MPH, Carrie Goodman5, MPH, George Hripcsak1, MD, MS.

1Department of Biomedical Informatics, Columbia University, 2School of Nursing, Columbia University, 3West Chester University, 4Georgia Institute of Technology, 5Clinical Directors Network

Introduction. In the United States alone, nearly 26 million adults and children have diabetes1. Self-management is a critical component of diabetes care2. Problem-solving, or one’s ability to remove barriers to engaging in self-management, is essential to successful diabetes self-management3, results in better diabetes self-care behaviors4 and leads to improvements in clinical outcomes5. There exist many innovative interventions for fostering individuals’ problem-solving skills. However, despite the high proliferation and popularity of electronic tools for diabetes self-management, few of them focus on discovery, learning, and problem-solving6,7. In this presentation, we will introduce Mobile Diabetes Detective (MoDD) – a novel electronic tool for engaging individuals with diabetes in problem-solving using data collected with self-monitoring. Using MoDD, individuals with type 2 diabetes identify blood glucose patterns that are above or below the ranges recommended by their healthcare providers, select behavioral triggers that contribute to these glycemic problems, set action-oriented goals for alternative healthy behaviors that can lead to improved glycemic control, and monitor their progress towards these goals. In this podium presentation we report the results of the qualitative study of user attitudes, preferences, and experiences with MoDD that emerged during their participation in a randomized controlled trial (RCT) of MoDD’s efficacy.

Methods. In this qualitative study, researchers conducted semi-structured interviews with individuals from economically disadvantaged communities and ethnic minorities who participate in a randomized controlled trial of MoDD, and examined MoDD usage logs. The transcripts of interviews were analyzed using inductive thematic analysis. MoDD usage logs were examined to determine usage patterns.

Results. Fifteen participants of MoDD RCT (out of 50 eligible) elected to take part in qualitative interviews. Usage log analysis showed that study participants logged into MoDD on average twice per week, reported 120 blood glucose readings during the 4 weeks of the study, and set 2 behavioral goals while in the trial. The qualitative interviews suggested that individuals used MoDD to make new discoveries about connections between different aspects of their daily activities (diet, exercise), and changes in their blood glucose levels, to experiment with new health behaviors, and to translate their discoveries into changes in self-management strategies.

Discussion. This study suggests that informatics interventions can play a proactive role in helping individuals to not only collect self-monitoring data and communicate with healthcare providers, but to also become active problem-solvers and to identify personally meaningful approaches to diabetes self-management through experimentation and discovery. In the relatively short four weeks of the study, MoDD participants were able to uncover new connections between their daily activities and changes in their blood glucose levels, experiment with new behaviors, set personally meaningful goals, and even see change in their blood glucose readings. However, the study also suggested several new directions for the design of informatics interventions for problem-solving and discovery. First, incorporating richer abilities for self-monitoring of daily activities can help applications like MoDD make tailored and personalized recommendations for problematic behaviors and suggest appropriate behavioral goals. Second, more explicit focus on barriers could help individuals to transfer their knowledge between different scenarios and situations. Finally, the study suggested the need to enrich theoretical landscape for informatics interventions for chronic disease self-management with theories that directly address the processes of discovery, experiential learning, and problem-solving.
References

![Figure 1: MoDD user interface; blood glucose readings log](image1)

![Figure 2: MoDD user interface; behavioral triggers and action-oriented goals](image2)
Protein Drug Target Prioritization for Illumination
Subramani Mani MBBS PhD, Daniel Cannon MS, Tudor Oprea MD PhD, Stephen Mathias PhD, Oleg Ursu PhD, Cristian Bologa PhD
University of New Mexico Health Sciences Center, Albuquerque, NM, 87131

Developing informatics methods for protein drug target prioritization to facilitate further illumination of these targets by knockout experiments on model organisms to study their phenotypic effects will advance discovery of novel drug targets for common and rare diseases. In this study we propose an informatics approach based on machine learning for prioritizing targets from the class of proteins labeled Tdark because not much is known about them based on our current state of knowledge.

Introduction and background: Out of the twenty thousand plus proteins in the expertly curated UniProt database [1] constituting the human proteome a subset of about 1800 constitute the four protein super families—G protein coupled receptors (GPCRs), Kinases, Ion channels (ICs) and Nuclear receptors (NRs) which are of special interest from a druggability perspective. Based on the current state of knowledge of these four family targets they have been categorized into six target development levels (TDLs)—Tclin+, Tclin, Tchem, Tmacro, Tgray and Tdark. These TDLs form a spectrum with Tclin+ targets having an approved drug with a known mechanism of action while Tdark targets have minimal information (<= 1 Gene Reference Into Function (RIF), <=38 antibodies and a low PubMed text mining score.) Researchers and pharmaceutical companies have typically focused on a small part of the genome-encoded proteome for drug targeting thus shedding light on these regions from a druggability perspective while leaving large tracts of the proteome dark. Based on their review of literature Hopkins et al. report that 399 non-redundant molecular targets have been shown to bind efficaciously with small molecules [2] out of more than 10,000 likely such targets in the human genome based on projections of ligand binding domains [3]. The challenge is to identify the relevant subset of the potential druggable targets that are represented in disease-linked proteins. Note that approximately ninety percent of the currently marketed drugs interact with the various proteins encoded by the genes represented in the exome [4]. This study is part of our larger goal to develop an in-silico framework for illuminating the druggable genome (IDG) using the resources of the IDG knowledge management center (KMC) and the target central resource database (TCRD) being developed as part of the IDG effort. The TCRD contains information about human protein targets with a focus on the four protein families of GPCRs, Kinases, ICs and NRs. In this work we focus on prioritizing the Tdark targets based on their similarity with Tclin+ targets using an informatics approach to begin the process of illuminating them from a druggability potential perspective.

Methods: We extracted various types of protein descriptors that were not used in defining the TDLs. Specifically we used the Gene Ontology (GO) terms, Protein Domains (PD), Expression (tissue expression profile) and Pathway representation for these targets as attributes to identify the Tdark targets that are similar to the Tclin+ class using a novel machine learning (ML) approach based on the confusion or misclassification matrices generated by a selection of tree learning algorithms. After preprocessing we had 183 GO attributes, 113 PD attributes, 59 expression attributes and 97 pathway attributes. Using these four databases we created four lists of Tdark proteins that were similar to Tclin+. Only the list of the best performing model (Random Forest) was used to create the lists. Six pairwise intersection sets generated by independent set of predictors (databases) were used to pick the prioritized Tdark targets. Using this derived similarity measure we ranked the top Tdark targets based on their luminosity or the potential for illumination. A list based on individual databases was also generated

Results: The Tdark targets prioritized based on being similar in at least two of the matrices generated by independent sets of predictors, that is, in the intersection of the list of Tdark targets classified as Tclin+ using two different datasets is shown in Table 1. Figure 1 shows the same prioritized targets based on their representation in the specific intersection set and also the probability of being similar to Tclin+ based on our approach. Out of the nine prioritized targets the majority (6) belong to the IC family, two are Kinases and one is a GPCR.

Discussion and Conclusion: We have proposed an informatics approach based on machine learning to prioritize targets from the category of Tdark about which not much is currently known regarding their potential as drug targets. Even though ICs are a comparatively smaller family with 342 targets we could prioritize six out of the 29 Tdark ICs because 151 of the ICs were in Tclin+. On the other hand we were able to prioritize only one out of the 454 Tdark GPCRs because only 93 of the 827 GPCRs are in Tclin+. We propose to prioritize additional Tdark targets using a ML approach making use of new databases as and when they are incorporated into the TCRD. The prioritized targets will be made available to researchers performing mouse and zebra fish knock out experiments to
further illuminate these potential targets. This study opens up the possibility of drug target prioritization using informatics methods and databases to propose new drug targets from the Tdark category for further illumination.

Table 1: The top prioritized Tdark target proteins. HGNC.Sym: HUGO Gene Nomenclature Committee official gene symbols. See http://www.genenames.org/

<table>
<thead>
<tr>
<th>HGNC.Sym</th>
<th>IDG.Family</th>
<th>UniProt/NCBI Gene database name</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAN3</td>
<td>Kinase</td>
<td>PAN3 poly(A) specific ribonuclease subunit homolog (S. cerevisiae)</td>
</tr>
<tr>
<td>TTYH3</td>
<td>IC</td>
<td>tweety family member</td>
</tr>
<tr>
<td>CNGA4</td>
<td>IC</td>
<td>cyclic nucleotide gated channel alpha</td>
</tr>
<tr>
<td>CSNK2A3</td>
<td>Kinase</td>
<td>casein kinase 2, alpha 3 polypeptide</td>
</tr>
<tr>
<td>GLRA4</td>
<td>IC</td>
<td>glycine receptor, alpha</td>
</tr>
<tr>
<td>GPR156</td>
<td>GPCR</td>
<td>G protein-coupled receptor 156</td>
</tr>
<tr>
<td>GPR98A</td>
<td>IC</td>
<td>Golgi pH regulator A</td>
</tr>
<tr>
<td>KCNK7</td>
<td>IC</td>
<td>potassium channel, subfamily K, member 7</td>
</tr>
<tr>
<td>KCNK15</td>
<td>IC</td>
<td>potassium channel, subfamily K, member 15</td>
</tr>
</tbody>
</table>

Figure 1: Prioritized Tdark targets for illumination

Intra-cluster correlation estimates for design of cluster-randomized trials and multi-clinic studies that utilize electronic health record data

Miguel Marino, PhD¹, John Heintzman, MD¹, Eve Dexter, MS¹, Stuart Cowburn, MPH², Jean P. O’Malley, MPH¹, Steffani R. Bailey, PhD¹, Rachel Gold, PhD MPH¹, Jennifer E. DeVoe, MD, DPhil¹,²

¹Oregon Health and Science University, Portland, OR; ²OCHIN Inc., Portland, OR

Introduction

As the adoption of electronic health records (EHRs) in the U.S. health care system rapidly increases, EHRs could provide a robust source of relevant data for cluster-randomized trials and multi-clinic observational studies. EHR data have the potential to improve the measurement of health outcomes and confounders, and enhance the ability to perform longitudinal assessments. In studies where group randomization is at the cluster level or patients are nested within a clinic and an individual-level analysis is to be performed, the classical assumption of independence between observations do not hold because subjects within a cluster are likely to have similar characteristics/outcomes. Standard approaches to sample size equations that ignore clustering typically are underpowered. Additionally, statistical models that ignore clustering typically exhibit narrower standard error estimates. To account for the relatedness of subjects within a cluster, sample size and power formulas use a measure of dependency between the subjects in a cluster known as the intra-cluster correlation (ICC) to provide correct estimates. Efficient design of these studies requires reliable estimates of the ICC of outcomes for subjects nested within clinics to use in sample size and power calculations. This study provides ICCs estimates of common outcomes found in EHR systems.

Methods

We used data from 31 Oregon primary care clinics and community health centers that adopted the networked EHR by 1/1/2010. The clinics are members of OCHIN, Inc., a non-profit organization that provides a fully hosted instance of Epic Systems’ practice management system (PMS) and electronic health record (EHR) to safety net clinics. PMS and EHR data are managed centrally at OCHIN, including regular validation and cleaning. Data for these analyses were extracted at OCHIN. The study population includes patients aged 19-64 who had ≥1 encounter at any of the study clinics between 1/1/2010-12/31/2011. We used generalized linear mixed effects models and cluster bootstrapping to construct intra-cluster correlation estimates and effective sample sizes for a large set of healthcare outcomes commonly used in primary care research, including body mass index, vital signs, encounter counts, chronic conditions, and preventive service utilization. Estimates of ICC close to 0 denote little correlation between subjects in the same cluster driving the effective sample size closer to the total number of subjects in the EHR study sample. As the ICC approaches 1, the responses from subjects in the same cluster are close to identical. This shrinks the effective sample size towards to number of clinics in the EHR sample.

Results

Study sample included 94,961 patients served by 495 physicians in 31 clinics. ICCs varied around a median value of 0.005 (IQR:0.001-0.021). Some characteristics (such as encounters per patient) had moderate correlation within clinic (ICC=0.156; 95% CI:0.101-0.233). Among clinical outcomes shown in Figure 1, we observe that physical characteristics varied the most with the highest ICC estimated at 0.055 for diastolic blood pressure and the lowest being the proportion of obese patients at an ICC close to 0. Prevalence of chronic conditions varied much more across patients within a clinic (ICC≤0.001). Preventive service uptake also showed little correlation within clinics (ICC≤0.010). Figure 2 shows the effective sample sizes obtained from the ICC estimates for various outcomes. There is a large variation in the effective sample sizes with the highest in comorbidity outcomes and a large reduction in the effective sample size for encounter counts.

Discussion

ICC estimates in this EHR data system are small in magnitude for the outcomes examined here. However, the estimated ICCs could affect the effective sample size required for a cluster-randomized trial or multi-clinic observational studies. The ICCs presented will be useful to help plan sample size and power calculations of future studies that utilize EHR data.
**Figure 1.** Estimates of ICC for Physical Characteristics, Delivery of Preventive Services and Comorbidity Outcomes.

**Figure 2.** Effective Sample Sizes Based on Observed ICC estimates and n=94,961.
Identifying Abnormal Anatomy on Temporal Bone Computed Tomography Reports Using Readily Available Natural Language Processing Software

Aaron J. Masino, PhD¹, Robert W. Grundmeier, MD¹, Jeffrey W Pennington, BS¹, E. Bryan Crenshaw III, PhD¹
¹The Children’s Hospital of Philadelphia, Philadelphia, PA, USA

Introduction

Electronic databases of health information have the potential to advance clinical research. However, significant amounts of valuable information remain locked in free text and for large databases are only accessible using natural language processing (NLP) methods. Most clinical researchers lack the expertise to use these methods on their own and must either collaborate with an established NLP expert or resort to manual chart review. Fortunately, NLP methods are increasingly packaged in well-documented toolkits, which may expand the ability of the clinical research workforce to extract useful information from free text. Audiology research represents one field of study where important anatomy information is available only in narrative reports from imaging studies such as high resolution CT scans of the temporal bones. In the AudGenDB database at The Children’s Hospital of Philadelphia there are over 15,000 unlabeled radiology reports, which represents a valuable resource for hearing loss research. Currently, researchers using this database can only use a keyword search to identify reports with specific characteristics (e.g. those that describe an abnormality). This is highly non-specific because nearly all reports mention normal and abnormal anatomical structures. Our objective was to provide a more robust search capability to identify reports with abnormalities using freely available python-based NLP and machine learning tools.

Methods

We used Python SciKit-Learn¹ and the Python Natural Language Toolkit² to dichotomously classify transcribed radiology reports as describing either normal or abnormal anatomy of ear-related structures. Any abnormality of the cochlea, cochlear nerve, vestibular aqueduct, semicircular canal, mastoid, tympanic membrane, ossicles, stapes, incus, malleus, scutum, or external auditory canal was considered a positive report. A set of 252 hand labeled CT scans of temporal bones were available for study (192 abnormal reports and 60 normal reports). Using stratified sampling we reserved 50 (20%) of the reports as a test set (38 abnormal and 12 normal reports). With the remaining 202 samples (the training set), we used 5-fold cross validation to train several models (e.g. logistic regression, support vector machines and decision trees), and select model parameters (e.g. regularization parameters) that maximized classification accuracy. Text pre-processing included word stemming and English stop word removal. Individual characters, character pairs and triples, words, and word pairs (bigramps) and triples (trigrams) were all considered as potential features. Both binary feature weighting (presence or absence of the feature) and feature counts were evaluated in model training and cross validation. Using the optimal features and parameters from the cross validation, we trained each model with all 202 training samples and evaluated performance on the 50 held-out test documents.

Results

The best performing feature sets, model parameters and predictive performance statistics on the test set are shown (Tables 1 and 2). A support vector machine with a linear kernel, regularization parameter=1.0, and character trigrams as input features achieved the best performance (recall=1.0, precision=0.95, F1 score=0.97). Naïve Bayes performed second best using character trigram occurrence as input features (recall=0.95, precision=0.95, F1 score=0.95). As a baseline, we performed a keyword search using the ear anatomy terms listed above. A document containing any of the keywords was labeled positive. The machine learning classification results are a significant improvement over the keyword search results (recall=1.0, precision=0.76, F1 score=0.86).

Discussion

Using a readily available open source NLP and machine learning toolkit and a modest number of labeled documents we were able to classify ear anatomy as normal vs. abnormal with both good recall and precision in temporal bone CT scan reports. The best-performing input features (character trigrams) and model can be constructed directly using the SciKit-Learn library. Based on these findings we conclude that researchers may be able to use basic NLP approaches readily available in open source toolkits to extract information in selected domains for their projects.
Table 1. Document features and parameters that achieved best performance in 3-fold cross validation for each model type.

<table>
<thead>
<tr>
<th>Model</th>
<th>Best Features Set</th>
<th>Best Model Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>Character trigram occurrence (binary)</td>
<td>N/A</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Character trigram counts</td>
<td>Inverse regularization parameter: 0.1</td>
</tr>
<tr>
<td>SVM (Linear)</td>
<td>Character trigram counts</td>
<td>Inverse regularization parameter: 1.0</td>
</tr>
<tr>
<td>SVM (Gaussian)</td>
<td>Character bigram occurrence (binary)</td>
<td>Inverse regularization parameter: 3.33, Kernel coefficient: 0.01</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>Character bigram counts</td>
<td>Maximum Depth: 2</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Character trigram occurrence (binary)</td>
<td>Maximum Depth: 5, Tree count: 150</td>
</tr>
</tbody>
</table>

Table 2. Predictive performance for each model with best cross-validation parameters on the test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F1 Score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.92</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.92</td>
<td>0.95</td>
<td>0.90</td>
<td>1.0</td>
</tr>
<tr>
<td>SVM (Linear)</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
<td>1.0</td>
</tr>
<tr>
<td>SVM (Gaussian)</td>
<td>0.90</td>
<td>0.94</td>
<td>0.88</td>
<td>1.0</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.86</td>
<td>0.92</td>
<td>0.84</td>
<td>1.0</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.88</td>
<td>0.93</td>
<td>0.86</td>
<td>1.0</td>
</tr>
<tr>
<td>Baseline (keyword search)</td>
<td>0.76</td>
<td>0.86</td>
<td>0.76</td>
<td>1.0</td>
</tr>
</tbody>
</table>

References

Developing InSPECt: An Interactive Surveillance Portal for Evaluating Clinical Decision Support

Allison B. McCoy, PhD1,2, Eric J. Thomas, MD3, Marie Krousel-Wood, MD, MSPH1,2, Susan C. Guerrero3, Rueben J. Applegate3, Dean F. Sittig, PhD3
1Tulane University, New Orleans, LA; 2Ochsner Health System, New Orleans, LA; 3The University of Texas Health Science Center at Houston, Houston, TX

Introduction
Healthcare providers are adopting electronic health records (EHRs) with clinical decision support (CDS) to improve patient safety and meet Meaningful Use requirements. Computerized alerts that prompt clinicians about drug-allergy, drug-drug, and drug-disease warnings or provide dosing guidance are most commonly implemented. Alert overrides can hinder improved patient outcomes, but providers are often overwhelmed with clinically insignificant alerts. Detailed evaluation of the appropriateness of alerts and clinician responses is necessary. CDS surveillance systems may facilitate evaluations and improve the appropriateness of alerts and responses.1,2

System Description
We developed InSPECt (Interactive Surveillance Portal for Evaluating Clinical Decision Support), a web-based dashboard that facilitates the review of CDS alerts and responses based on a previously implemented condition-specific dashboard that allowed clinical pharmacists and informatics personnel to review CDS alert responses in the context of patients at high risk for ADEs.3 This podium presentation will include a demonstration of InSPECt and its ability to facilitate evaluations of the appropriateness of a set of alerts or an individual alert and responses.

InSPECt was implemented using the JavaServer Faces (JSF) Java Specification and is backed by a Microsoft SQL Server database. It uses the jQuery Javascript library to enable interactivity, including the jqPlot jQuery plugin to generate graphs. Alert log and clinical data are retrieved weekly by scheduled, automated tasks that query the clinical data warehouse and transform the information to a generic data model for alert evaluation to facilitate the inclusion of alerts from additional EHRs in the future. The data model includes the alerts (unique identifier, patient, ordering provider, date/time, alert type, ordered item identifier, triggering item identifier, override flag and reason, and displayed text), triggering items, patients, and providers, where ordered and triggering items may include medications, laboratory orders or results, clinical problems, or allergies. InSPECt exists behind the institutional firewall and uses LDAP authentication with a manual approval process for allowed users. All user interactions are logged in the application database for further usability and security analyses.

InSPECt consists of two view types: the alert and the patient detail views. The alert detail view displays a graph of alert rates over time, including total and overridden alerts, along with a list of all logged alert instances. It allows reviewers to identify inappropriate alerts, showing details such as alert time; triggering medication(s), lab value, or allergy; patient demographics; and clinician name and service. The display can be filtered or sorted on any column (Figure 1). The patient detail view displays a graph of alerts and ordered medications over time, along with a detailed timeline for a patient in reverse chronological order to give context for the alerts. The timeline includes all orders, problems, lab results, and alert interactions occurring in the patient’s EHR and may be sorted on any column. Reviewers can use the patient detail view to understand clinician actions and patient condition changes occurring in conjunction with alert overrides without having to redirect to and search a patient’s EHR (Figure 2).

Conclusions
Development of an EHR-independent, interactive web-based dashboard for CDS alerts is feasible and may facilitate alert evaluations and improvements. Further research is warranted and underway to evaluate the dashboard for usability and ability to improve alert appropriateness and resulting patient safety.

Acknowledgments
This work was supported by NLM Grant 1K22LM011430-01A1, a UTHHealth Young Clinical and Translational Sciences Investigator Award (KL2 TR 000370-06A1), and NCRR Grant 3UL1RR024148.
References


Robust Sentence Segmentation for Clinical Text

Timothy A Miller, PhD, Sean Finan, Dmitriy Dligach, PhD, and Guergana Savova, PhD
Boston Children’s Hospital Informatics Program, Harvard Medical School, Boston, MA

Introduction:
Natural language processing (NLP) has been applied to clinical text for a variety of use cases, including building virtual cohorts or retrospective research.\textsuperscript{1,2,3} Machine learned models for several core NLP tasks, including detecting sentence boundaries, are part of Apache cTAKES, an open source clinical NLP tool (ctakes.apache.org). These models are designed to be portable, but in practice new domains can affect performance.

Sentences are crucial linguistic units (e.g., for extracting relations between clinical entities) and accurately finding them is important for downstream processing. Existing solutions to this task do not solve the problem in a way that addresses the diversity of clinical text. While methods from the general domain work on biological text,\textsuperscript{4} both of those domains have much more uniformity of editing and formatting than clinical text. The cTAKES system description paper\textsuperscript{5} describes excellent results on clinical text but does not report results on outside datasets. Recent work on newspaper text\textsuperscript{6} described high quality features, but ignored formatting and relies heavily on token features.

The major problem presented by clinical text is inconsistent formatting. Inconsistent use of newlines is especially problematic on the largest freely available corpus, MIMIC\textsuperscript{3}. Most segmenters ignore newlines, while cTAKES recognizes that in clinical text, a newline character often ends a sentence (in lieu of a period). However, newlines in the MIMIC corpus are highly ambiguous, sometimes inside sentences while other times ending them.

Methods:
We generalize the problem to one of character sequence tagging. Previous approaches used discriminative classifiers on a small set of candidate sentence-splitting characters. However, in clinical data a sentence break could be indicated by something as subtle as a series of spaces. Classifying the sequence of characters as the Beginning, Inside, or Outside of a sentence allows us to use similar features as previous systems while allowing arbitrary sentence boundaries. We avoid major performance penalties by only classifying non-alphanumeric characters.

We also introduce features intended to abstract away from word identity. Common words are probably valuable, so we keep some word features, but we will be unlikely to obtain a gold standard dataset large enough to capture words that are useful in all clinical sub-domains. We thus experiment with character, shape, and position features.

Character features are applied to the current character being classified as well as a window around it of n=3. These features include identity, upper or lower case, and whether it is a digit, whitespace, or punctuation. These are meant to capture patterns like punctuation around the current location (which may indicate an abbreviation), or major paragraph breaks (multiple newlines). Shape features map every character in a word to its character class and then convert it to a regular-expression-like representation. For example, capitalized words have a representation like “Ul*” to indicate an upper case character followed by a sequence of lower case characters. Shape features are applied to the two tokens surrounding the current character. Position features compare the position of a newline character being classified on a line of text to document-level line lengths. We compare the position of a newline on its line to a histogram of counts of line lengths and check whether the next word extends the line too far.

These features augment a set of widely used features from Gillick\textsuperscript{7} and OpenNLP (opennlp.apache.org): Left and right token identity, token length, right token capitalization, a log count of times the left token occurred without a period, log count of times the right token appeared lower-cased, and combined left token and right capitalization.

To evaluate these methods we manually annotated a set of MIMIC notes, including a subset of the discharge summaries, nursing notes, and radiology reports from patients numbered 1-40, consisting of 3245 sentences. We implemented the features in ClearTK (cleartk.org), a UIMA-based machine learning wrapper, using a support vector machine classifier via the LIBLINEAR library. We performed a 5-fold cross-validation to make optimal use of the relatively small dataset. We experimented with several different conditions, shown in column headers of Table 1. Performance is evaluated using precision (p), the percent of predicted spans that had exact matches in the gold standard, recall (r), the percent of gold standard spans that were found, and F1=2*p*r / (p+r).

Results:
Results are in Table 1. The cTAKES baseline performs poorly on this dataset. Gillick features in a new model trained on new data are a major improvement but still score low. Character features improve performance the most of any feature type. Shape features are less valuable on their own while position features are the least valuable on
their own. Combining all features does not improve performance over character features alone. Pairing character features with position has nominally the best result, but is essentially equivalent to character features alone.

**Discussion:**

Even with major improvement, the models trained and tested on MIMIC data do not attain performance comparable to the general domain (Gillick’s best result has an error rate of less than 1%). This may be partially addressed with more annotated data, though more diverse data sources may actually hurt performance. The sentence segmentation problem is probably inherently harder on unedited text than on edited newspaper or biomedical journal text.

The best configuration and models resulting from this work will be included as part of Apache cTAKES.

**Table 1.** “cTAKES baseline” is the off-the-shelf performance of cTAKES sentence segments, “Gillick features” is a system trained on this data with features from Gillick (2009). Columns starting with “+” indicate individual or paired feature types added to Gillick features, and “Combined features” is a system with all new feature types in addition to Gillick features.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>cTAKES Baseline</th>
<th>Gillick features</th>
<th>+Character features</th>
<th>+Shape features</th>
<th>+Position features</th>
<th>+Char +Shape</th>
<th>+Char +Position</th>
<th>Combined features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.244</td>
<td>0.663</td>
<td>0.863</td>
<td>0.739</td>
<td>0.674</td>
<td>0.861</td>
<td><strong>0.865</strong></td>
<td>0.859</td>
</tr>
<tr>
<td>Recall</td>
<td>0.458</td>
<td>0.463</td>
<td>0.781</td>
<td>0.569</td>
<td>0.501</td>
<td>0.778</td>
<td><strong>0.785</strong></td>
<td>0.778</td>
</tr>
<tr>
<td>F1</td>
<td>0.318</td>
<td>0.534</td>
<td>0.820</td>
<td>0.638</td>
<td>0.575</td>
<td>0.817</td>
<td><strong>0.823</strong></td>
<td>0.817</td>
</tr>
</tbody>
</table>

**Acknowledgements:**

This work was supported by the following grants: 1R01GM103859-01A1, 1U24CA184407-01, and R01 R01GM090187.

**References**


7. Mohammed Saeed, MD, PhD, Mauricio Villarroel, MBA, Andrew T. Reisner, MD, Gari Clifford, PhD, Li-Wei Lehman, PhD, George Moody, Thomas Heldt, PhD, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark, MD, PhD. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database 2011; 39:952-960.
An Analysis of Patient Portal Use in the Acute Care Setting

Eli Mlaver¹, Anuj K. Dalal, MD¹,², Harry Reyes Nieva¹,², Frank Chang, MSE¹, John Hanna¹, Sucheta Ravindran, MS¹, Kelly McNally¹, Diana Stade¹, Constance Morrison¹, David Bates, MD, MSc¹,², and Patricia Dykes, RN, PhD¹,²

¹Brigham and Women’s Hospital, Boston, MA, ²Harvard Medical School, Boston, MA

Introduction: Utilizing health information technology to engage patients in their healthcare is a growing trend, yet considerable gaps in knowledge remain regarding best practices for implementation in acute care settings.¹ Our team implemented a web-based patient-centered toolkit (PCTK), accessible on bed-side iPads, in the acute oncology ward as part of the PROSPECT (Promoting Respect and Ongoing Safety through Patient-centeredness, Engagement, Communication and Technology) project at Brigham and Women’s Hospital. The PCTK provided patients with a breadth of information related to their hospitalization. The portal’s content and design were developed using a focus group- and interview-based participatory design approach.² Given the novelty of such a program, there is still much work to be done to identify standardized techniques to gauge use of various features. The aims of this analysis were to identify consistency of and trends in patient use by length of stay in order to strengthen our understanding and guide future development of this and similar patient engagement technologies.

Methods: We reviewed PCTK activity from July 15, 2014 to January 13, 2015 using time-stamped clicks. We stratified patients into three groups based on length of stay. We excluded clicks made within an hour of the consent and enrollment period of each user to account for the tutorial we provided to users. We also calculated consistency of PCTK use as a percentage of total days used over the course of the care unit stay. We then identified changes in use of PCTK features among the short, medium, and long-stay groups. We compared continuous variables using ANOVA and categorical variables using the Chi-squared test, and we considered a p-value of less than 0.05 statistically significant.

Results: Of 137 users, 43, 51, and 43 had a length of stay of 1-4 days, 5-14 days, and 15+ days, respectively, with no significant demographic differences among the three groups (Table 1). Ninety-nine (72%) continued to use the portal after the enrollment period, and these users on average accessed the PCTK on 63% of the days they were hospitalized. We identified a total of 13,771 post-enrollment clicks based on time-stamp data, which were categorized into twelve groups (Table 2). Goals/Plan of Care, Lab Test Results, and Medications were clicked most among all groups, and all three features garnered a higher percentage of clicks among the longer-stay groups. There was a decrease in percentage of clicks in the longer-stay groups for safety information and the messaging system features—the fourth and fifth-most clicked features, respectively.

Discussion: We observed that patients who used the portal after the enrollment period had high retention of use regardless of length of stay. The data also suggest that oncology patients with longer hospitalizations seem to engage with their plan of care, test results and medication information. Further analysis with this and other patient populations should clarify whether these observations are generalizable and will the impact of other variables such as diagnosis and medication timing as well as other co-morbidities. This should provide insight into how to customize the availability of portal features by clinical unit, service, or patient. Further analysis will also look into hourly rather than daily portal use to see if patients regularly use the portal multiple times per day or at patterned times of day, which could help focus clinicians as to when to expect questions and concerns to be voiced. Insights gleaned from this analysis have directed improvement in the development of the PCTK, and we hope that it will also guide other institutions that are developing or providing similar programs.

Acknowledgements: The Brigham and Women’s Hospital PROSPECT project is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation.
Table 1. Patient demographics by length of stay in the acute oncology unit

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Overall (N = 137)</th>
<th>1-4 Days (N = 43)</th>
<th>5-14 Days (N = 51)</th>
<th>15+ Days (N = 43)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (+ Standard Deviation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, years</td>
<td>57.42 (14.88)</td>
<td>61.33 (16)</td>
<td>58.59 (13.35)</td>
<td>52.12 (14.45)</td>
<td>0.012</td>
</tr>
<tr>
<td>Length of stay, days</td>
<td>10.78 (9.61)</td>
<td>2.70 (1.06)</td>
<td>8.04 (2.67)</td>
<td>22.12 (9.03)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.531</td>
</tr>
<tr>
<td>Women</td>
<td>86 (63)</td>
<td>28 (65)</td>
<td>29 (57)</td>
<td>29 (67)</td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>51 (37)</td>
<td>15 (35)</td>
<td>22 (43)</td>
<td>14 (33)</td>
<td></td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.400</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>121 (88.32)</td>
<td>42 (97.67)</td>
<td>43 (84.31)</td>
<td>36 (83.72)</td>
<td></td>
</tr>
<tr>
<td>Black/African American</td>
<td>4 (2.92)</td>
<td>0 (0)</td>
<td>3 (5.88)</td>
<td>1 (2.33)</td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>3 (2.19)</td>
<td>0 (0)</td>
<td>2 (3.92)</td>
<td>1 (2.33)</td>
<td></td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>3 (2.19)</td>
<td>0 (0)</td>
<td>1 (1.96)</td>
<td>2 (4.65)</td>
<td></td>
</tr>
<tr>
<td>Other/Unknown</td>
<td>6 (4.38)</td>
<td>1 (2.33)</td>
<td>2 (3.92)</td>
<td>3 (6.98)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. PCTK usage by length of stay in the acute oncology unit

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Overall (N = 99)</th>
<th>1-4 Days (N = 29)</th>
<th>5-14 Days (N = 33)</th>
<th>15+ Days (N = 37)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (+ Standard Deviation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistency, percentage²</td>
<td>63.49 (42.28)</td>
<td>84.77 (49.61)</td>
<td>40.26 (30.27)</td>
<td>31.90 (27.22)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>PCTK Feature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Goals/Plan of Care</td>
<td>2815 (24.50)</td>
<td>425 (23.06)</td>
<td>703 (22.65)</td>
<td>1687 (25.79)</td>
<td></td>
</tr>
<tr>
<td>Clinical Problems</td>
<td>322 (2.80)</td>
<td>72 (3.91)</td>
<td>117 (3.77)</td>
<td>133 (2.03)</td>
<td></td>
</tr>
<tr>
<td>Schedule</td>
<td>160 (1.39)</td>
<td>33 (1.79)</td>
<td>48 (1.55)</td>
<td>79 (1.21)</td>
<td></td>
</tr>
<tr>
<td>Care Team</td>
<td>575 (5.00)</td>
<td>87 (4.72)</td>
<td>162 (5.22)</td>
<td>326 (4.98)</td>
<td></td>
</tr>
<tr>
<td>Messaging</td>
<td>500 (4.35)</td>
<td>93 (5.05)</td>
<td>159 (5.12)</td>
<td>248 (3.79)</td>
<td></td>
</tr>
<tr>
<td>Safety</td>
<td>726 (6.32)</td>
<td>176 (9.55)</td>
<td>219 (7.06)</td>
<td>331 (5.06)</td>
<td></td>
</tr>
<tr>
<td>Lab Test Results</td>
<td>1394 (12.13)</td>
<td>181 (9.82)</td>
<td>374 (12.05)</td>
<td>839 (12.82)</td>
<td></td>
</tr>
<tr>
<td>Medications</td>
<td>967 (8.42)</td>
<td>142 (7.7)</td>
<td>270 (8.7)</td>
<td>555 (8.48)</td>
<td></td>
</tr>
<tr>
<td>Food and Diet</td>
<td>309 (2.69)</td>
<td>66 (3.58)</td>
<td>89 (2.87)</td>
<td>154 (2.35)</td>
<td></td>
</tr>
<tr>
<td>Discharge Checklist</td>
<td>220 (1.91)</td>
<td>63 (3.42)</td>
<td>68 (2.19)</td>
<td>89 (1.36)</td>
<td></td>
</tr>
<tr>
<td>PCTK User Guide</td>
<td>158 (1.38)</td>
<td>29 (1.57)</td>
<td>56 (1.8)</td>
<td>73 (1.12)</td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td>159 (1.38)</td>
<td>15 (0.81)</td>
<td>57 (1.84)</td>
<td>87 (1.33)</td>
<td></td>
</tr>
<tr>
<td>Home/Notifications³</td>
<td>3184 (27.71)</td>
<td>461 (25.01)</td>
<td>782 (25.19)</td>
<td>1941 (29.67)</td>
<td></td>
</tr>
</tbody>
</table>

¹Tutorial clicks occurring within the first hour of enrollment were excluded.
²Consistency was measured as number of days PCTK was used divided by length of stay.
³Clicks in this category correspond to general use of PCTK rather than any distinct feature.

References

Translating Electronic Clinical Quality Measures to Executable, Portable, and Customizable Workflows in KNIME

Huan Mo, MD, MS; Jennifer A. Pacheco; Richard Kiefer; Luke V. Rasmussen; Jyotishman Pathak, PhD; Joshua C. Denny, MD, MS; William K. Thompson, PhD
Vanderbilt University, Nashville, TN; Northwestern University, Chicago, IL; Mayo Clinic, Rochester, MN; NorthShore University HealthSystem, Evanston, IL

Introduction

The widespread adoption of Electronic Health Records (EHRs) has led to the development of a range of electronic clinical quality measures (eCQMs), as well as EHR-driven phenotype algorithms for clinical research. Such quality measures and phenotype algorithms are typically developed at a single site and then ported across multiple sites. To facilitate this process, it is necessary to adopt an unambiguous formal language for algorithm representation, such as the Office of the National Coordinator (ONC)-sponsored Quality Data Model (QDM). However, there is currently no widely available mechanism for translating QDM-based specifications into executable workflows that can operate on local EHR data, making the porting process difficult, time-consuming, and error-prone. As part of the Phenotype Execution Modeling Architecture (PhEMA) project, we are addressing this gap through the creation of a translator from QDM-based specifications to executable workflows for use with the open-source Konstanz Information Miner (KNIME) data analytics platform. These workflows are easily shared across multiple sites, can be connected to local repositories with varying data structures, and allow for customization to account for heterogeneous local EHR data.

Methods

The input format to our translator is the JSON representation of QDM originally developed for Project Cypress, a tool for testing eCQM validity. The translation target is KNIME, which offers a user-friendly graphical user interface with workflows represented graphically. Graph nodes represent data transformations, and graph links represent data flow (Figure 1). Our QDM2KNIME translator has three parts. First, it provides a near-complete collection of reusable KNIME nodes based on standard QDM elements and attribute sets, covering data types, temporal relationships, logical operations, and aggregative functions. Second, it provides a Java API for the stepwise construction of KNIME workflows for different translators and web applications. This includes the creation and setup of KNIME nodes, connections between them, and other necessary parts of the workflow. Third, it provides functionality for converting Cypress patient test data into a tabular format to test eCQMs inside of KNIME. We used QDM2KNIME to convert 10 published eCQMs into KNIME workflows, and tested four of them (CMS30, CMS123, CMS126, CMS179) using the converted Cypress test patient data. In addition, two of the authors (HM, JAP), tested eCQM CMS30 on EHR patient data by translation from eCQM to KNIME and execution on two different institutional repositories (VU and NU).

Results

Running the Cypress test data through the test eCQMs produced the desired results: all test patients were scored correctly with respect to each of the four selected eCQMs. This result provides us with basic validation that QDM2KNIME is generating executable workflows that faithfully represent the logic of their source specifications. A more stringent test of feasibility is to use actual EHR patient data from multiple institutions. Measure CMS30 was translated into KNIME and executed at both VU and NU. Implementation required each to configure the data access and mapping workflow nodes (Figure 1). The resulting scores broadly matched our expectations, however further evaluation against gold standard EHR data sets is planned for future work.

Discussion

We have determined the feasibility of automatically translating QDM-based eCQMs into executable KNIME workflows that can be shared across institutions and appropriately customized to handle patient data from multiple, heterogeneous EHRs. Our next steps are to integrate the QDM2KNIME translation software with the full suite of PhEMA tools, which includes a QDM-based authoring tool and the PheKB repository of phenotype algorithms. This integrated tool suite will make it possible to author and share QDM-based eCQMs and algorithms across multiple institutions with far less effort than is currently required.
Figure 1: Implementation of the eCQM CMS30 (Statin at Discharge) on KNIME. This workflow was generated automatically from a QDM-based representation of the quality measure. Nodes in the workflow graph represent data sources and outputs, data transformations implementing application logic. Connections between nodes represent data flow. After auto-generation, workflows can be modified manually using the KNIME graphical user interface.

Acknowledgements

This work has been supported in part by funding from PhEMA (R01 GM105688) and eMERGE (U01 HG006379, U01 HG006378 and U01 HG006388).

References


Improving Radiology Procedure Identification for Inferior Vena Cava (IVC) Filters using EHR Text

Dalia Mobarek, MD 1,2, Najeebah Bade, MD 1,2, Benjamin Viernes, MPH 3,4
Patricia Nechodom, MPH 3,4, Frederick R. Rickles, MD 1,2, Scott Duvall, PhD 3,4
1 Veterans Affairs Medical Center, 2 School of Medicine, George Washington University, Washington DC, USA 3Veterans Affairs Medical Center, 4 School of Medicine, University of Utah, Salt Lake City, Utah, USA

Abstract
Insertion of inferior vena cava filters is an anticoagulation alternative in patients with venous thromboembolism. Insertions have increased without corresponding increase in retrieval despite FDA communications to remove filters as soon as possible to decrease complications. This retrospective study used VA national administrative data and found 8,314 IVC filter placements and only 980 retrievals. Placements and retrievals also occur without a standard CPT and/or ICD9 coding. We propose an alerting system to identify retained filters

Introduction
Venous thromboembolism (VTE), including deep venous thrombosis (DVT) and pulmonary embolism (PE), represents a major cause of preventable morbidity and mortality in the United States. The preferred standard initial therapy for VTE is anticoagulation; however, anticoagulation may be contraindicated or considered too risky to initiate in some. In this population, insertion of an inferior vena cava (IVC) filter has become a successful alternative. Filter placement has increased significantly over recent years, particularly since the introduction in the U.S. of retrievable filters in 2003. Filtering has been shown to offer short term protection against PE, but have been implicated in increased complication rates in the long term. A 2010 US FDA safety communication stated that retrievable IVC filters should be considered for removal as soon as protection from PE is no longer needed. Despite this, the actual rate of retrieval is surprisingly low. The aim of this study is to use a large patient database to define the cohort of patients in whom IVC filters have been placed throughout the VA system and study their outcomes.

Methodology
This is a retrospective cohort study using data available through Veteran’s Affairs Informatics and Computing Infrastructure (VINCI), which houses data from all VA medical facilities in the nation. It includes structured and unstructured data sources, administrative databases and the EHR. We identified patients with any CPT and ICD9 procedure codes related to placement or retrieval of IVC filters between 2006 and 2014. Before 2012, there were no unique standardized codes for IVC filter placement or retrieval, with multiple procedures using these non-unique codes. We searched EHR text of the cohort patients using key words to determine whether they had a filter placed or removed. Validation and improvement of the cohort is underway through keyword search and manual annotation. Once we establish the final cohort, we will report the number of retrieved IVC filters, duration of use, and complications. These methods for identifying filter procedures will be used to develop an EHR alert to track patients with retained filters.

Results
From 1/1/12 to 9/30/14, 3,553 IVC filters were placed. 698 subjects expired within 90 days of procedure. 331 filters were retrieved as determined by the CPT code for filter retrieval. Prior to 2012, 8,361 IVC filters were placed. The CPT code for retrieval used prior to 2012 was not exclusively used for IVC filters, therefore keyword search of the clinical notes in the EHR and manual chart annotation will be used to establish the number of retrieved IVC filters.

Discussion
IVC filter retrieval remains low throughout the VA medical system, despite the FDA safety communications. We anticipate finding multiple complications in the inappropriately retained IVC filter group; however, these results will be reported after establishment of the cohort. Based on preliminary work, we see a need for tracking retained filters utilizing text mining as a tool; further characterization of the latter is underway. This approach of enhancing the cohort through text searches and chart review contributes substantially to the production data more closely matching the clinical realities. This work is most significant as an example of how textual healthcare data can be incorporated into computational analysis, not solely to identify IVC retrieval but for other similar uncoded healthcare issues.

References
Computable Phenotypes enabled by the i2b2 Validation Platform

Shawn Murphy\(^1\), Vivian Gainer\(^1\), Victor Castro\(^1\), Alyssa Goodson\(^1\), Lori Phillips\(^1\), Sheng Yu\(^2\), Tianxi Cai\(^2\)

Massachusetts General Hospital and Partners Healthcare\(^1\), Harvard School of Public Health\(^2\), Boston, MA

Abstract: If Electronic Healthcare Record (EHR) Data is to be used for scientific discovery it must either contain parameters about its overall accuracy or we must invent methods to assess its accuracy and derive clinical meaning from it. Validation of raw EHR data accuracy often reveals poor sensitivity and positive predictive values. However, raw data consisting of codes and narrative text can be molded in most cases into accurate values. The technique for performing this transformation can be complex and requires a rigorous approach\(^1\). We wrapped a set of data transformation tools into the Informatics for Integrating Biology and the Bedside (i2b2) Phenotype Validation Platform. The platform provides both a methodology and tools and can scale to calculate large numbers of computed phenotypes and publish them back into the i2b2 database where they can be queried with the i2b2 query tool and extracted for research. Using the platform, eight computed phenotypes were created and validated in a systematic fashion. We are working to make the platform generally available for use in i2b2 and SHRINE installations.

Introduction: I2b2 is open source software that provides clinical investigators with the tools necessary to integrate and query qualitative and quantitative medical record and clinical research data\(^2\). Over 120 hospitals and academic medical centers have installed the i2b2 suite of software tools and have populated their databases with a broad range of EHR data. Unfortunately, much of the EHR data in its raw form is not useful for scientific applications because of its poor quality\(^3\). Fortunately, the existence of many types of rich data such as coded diagnoses, coded medications, and provider notes gives the scientist an opportunity to derive highly accurate data. Yet the process to derive accurate, high quality data from the raw EHR data is complex and unintuitive. Therefore, we created a platform to guide the process and perform the calculations. Methods: A workflow consisting of 6 steps was created and used to derive 8 new accurate phenotypes from the raw EHR data using tools on the Platform. In Step 1 we define the phenotype of interest, which could be as simple as a specific diagnosis but could also be more complex, such as a treatment outcome. The second step is to very loosely identify patients who may have the phenotype using easily available EHR data to allow further raw data to be collected with sufficient detail (such as narrative text) for calculations but also to allow appropriate signoffs from regulatory boards to use the detailed data for this limited set of patients. The third step is to take a sample of the patients from Step 2 and manually create a gold standard of those that have the required phenotype. The fourth step is to create a comprehensive list of features from the detailed data estimated to be useful to statistically define a phenotype. Automated tools in this step are especially useful for increasing efficiency. The fifth step is to arrange the features into a file to allow a statistical package to operate upon the features and fit the gold standard set of patients for the phenotype to the features. The sixth step is to validate the algorithm with patients outside of the gold standard training set. Results: We used the workflow of the Platform to transform eight phenotypes into new codes with high accuracy (AUC > 0.92). The phenotypes were: Bipolar Disease, Congestive Heart Failure, Coronary Artery Disease, Crohn’s Disease, Multiple Sclerosis, Rheumatoid Arthritis, Type 2 Diabetes, and Ulcerative Colitis. For each patient the new codes could now be associated with values such as sensitivity and positive predictive value from the “Confusion Matrix”. These values could then be used to query for the desired tradeoff of sensitivity/specificity when searching for patients with the desired phenotype. Discussion: The limitations of raw data from the EHR can disrupt approaches to scientific discovery using this data. We have been developing systematic approaches to curating this data to provide new derived codes of greater accuracy for phenotypes of interest. The tradeoff is that the process is complex and time-consuming for deriving these codes. To make these techniques more available to mainstream computing, we have encapsulated them into a platform that make the process easier to manage.
Figure 1: Eight new highly accurate codes for phenotypes were created from the platform described and integrated into i2b2 so that they could be queries and used for further analysis along with other data in the database.

Will they participate? Predicting patients’ response to clinical trial invitations

Yizhao Ni, PhD1, Andrew F. Beck, MD, MPH1, Regina Taylor, MA1, Jenna Dyas, BS1, Imre Solti, PhD, MD, MA, MSc1, Judith W Dexheimer, PhD1

1Cincinnati Children’s Hospital Medical Center (CCHMC), Cincinnati, OH

Introduction

Challenges with patient recruitment for clinical trials are recognized as a major barrier to the timely and efficacious conduct of translational research. Studies have identified various patient- and physician-level factors that impact the successful recruitment of patients for clinical trials1,2. This work has focused on factors impacting patients’ enthusiasm for contributing to a trial, including acuity of their clinical problems, demographics and socioeconomic status (SES), and trial characteristics2,3. To take the next step, our objective is to develop a machine-learning model to predict a patient’s response (i.e., if the patient agrees or declines) before he/she is approached for a clinical trial.

Data and Method

We focused on clinical trials for pediatric patients who visited the Emergency Department (ED) at Cincinnati Children’s Hospital Medical Center (CCHMC) between 1/1/2010 and 8/31/2012. All 18 ED clinical trials that required patients’ consent before enrollment were included in the study. During the study period patients in 3134 encounters were eligible for one of the trials and were approached for consent. Patients in 1911 encounters agreed to participate in the trials (consent rate 61%). Based on the literature, we retrospectively collected a list of variables from multiple sources (Table 1), including the patients’ visit data (e.g. demographics and acuity of clinical problems), a proxy of SES, and the characteristics of the clinical trials. To estimate their SES, the patients’ addresses were geocoded to identify the census tract in which they lived. Ten socioeconomic characteristics from the 2008-2012 US Census American Community Survey were then extracted as proxies for individual-level SES (Table 1).

Patients’ responses from the 3134 encounters served as a set of gold standard to train and evaluate the predictive models. We performed a stratified random sampling based on numbers of patients approached for each trial to split the data into two sets, 90% for training and development and 10% for evaluation. The baseline is a random response predictor using the binomial probability model, where the probabilities of agreeing/declining a trial invitation were optimized by maximum likelihood estimation. We then compared the baseline with logistic regression (LR), where ten-fold cross-validation was used to optimize the cost parameter (C) screened from 2^-20 to 2^20. The nominal variables (e.g. gender and insurance type) were converted to binary features, while the numerical variables (e.g. SES data) were discretized into ordinal features using the Chi-square test. Macro-average positive predictive value (PPV), recall (R), negative predictive value (NPV), specificity (SPC) and the area under the ROC curve (AUC) were calculated to assess the algorithm performance. In addition, we tested the three variable sets (visit data, SES data and trial characteristics) individually and in combination to validate their respective contribution to the predictive model.

Results and Discussion

The performances of the baseline and the LR algorithm are presented in Table 2. Compared with the baseline, the LR algorithm achieved better AUC. The LR with all variables achieved the best AUC (set 7). Improvements were statistically significant over the LR s using individual sets, and the combination of visit and SES data (set 1-4). Among the variable sets, the trial characteristics (set 3) were shown to be more informative than the others. The visit data was less informative, but combining it with the trial characteristics (set 5) significantly improved the AUC (p=1.4E-2 under paired t-test). Including SES data barely improved the performance, possibly due to the fact that the data is ecological, referring to the status of the geographic areas and not specific to individual households. Table 3 shows the features from the multivariable LR model that were associated with patients’ response at p=0.1 significance level. We observed that patients were more likely to participate in disease-specific and interventional studies. They were less likely to participate in multi-center trials, more invasive trials, and trials that required follow-up visits or phone calls. Multiple factors on patients’ visits were significantly associated with their response to a trial invitation, including the seasons of their visits, their arrival means, length of stay and disposition. Finally, some of the patients’ characteristics, such as race and SES, could also influence their enrollment preference.

The developed algorithm and the observations could have potential for significant impact in the strategic planning of clinical trial enrollment. The algorithm could facilitate recommendations of trials to patients in ways that maximize chance of enrollment. We could also recommend the patient to a trial that has higher priority in strategic planning and with an acceptable level of enrollment possibility.
Table 1. List of variables collected for the study.

<table>
<thead>
<tr>
<th>Patient Visit Data</th>
<th>Census Tract-Level SES Data</th>
<th>Clinical Trial Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics (age, gender, race, ethnicity)</td>
<td>Percentage (PCT) of persons at &lt;50% poverty line</td>
<td>Complexity of the trial (1-3, 1:simple; 3: complex)</td>
</tr>
<tr>
<td>Arrival means (e.g. walk in or by private car)</td>
<td>PCT of persons at &lt;100% of poverty line</td>
<td>Conductor of the trial (parents, CRGs, nurses/physicians)</td>
</tr>
<tr>
<td>Arrival time interval (e.g. 8am-10am, at two-hour increment)</td>
<td>PCT of housing units that are unoccupied</td>
<td>Study type (observational, interventional, other)</td>
</tr>
<tr>
<td>Arrival season (e.g. Spring)</td>
<td>PCT of households who rent their home</td>
<td>Amount of compensation</td>
</tr>
<tr>
<td>Acuity of the problem (1-5, 1:urgent; 5: not urgent)</td>
<td>PCT of households who do not own a car</td>
<td>Randomized trial Y/N</td>
</tr>
<tr>
<td>Guardian presence Y/N</td>
<td>Median household income</td>
<td>Multi-center trial Y/N</td>
</tr>
<tr>
<td>Length of stay (in hours)</td>
<td>Median value of owner-occupied house</td>
<td>Requiring samples Y/N</td>
</tr>
<tr>
<td>Disposition (admit/discharge/transfer)</td>
<td>PCT of households with ≥1 person/room</td>
<td>Requiring follow up visit Y/N</td>
</tr>
<tr>
<td>Insurance type (commercial/Medicare/self-pay)</td>
<td>PCT of persons aged ≥15 years who have never married</td>
<td>Only enrolling Medicare or self-pay patients Y/N</td>
</tr>
<tr>
<td></td>
<td>PCT of persons aged ≥25 years with &lt;12th-grade education</td>
<td>Involving sensitive topics Y/N (e.g. psychiatric disorders)</td>
</tr>
<tr>
<td></td>
<td>Unemployment rate for persons ≥16 years in the workforce</td>
<td>Invasiveness (1-5, 1: noninvasive; 5: highly invasive)</td>
</tr>
</tbody>
</table>

Table 2. Classification performance of the baseline and the logistic regression with different variable sets.

<table>
<thead>
<tr>
<th>Task</th>
<th>Ten-fold Cross Validation Performance [%]</th>
<th>Test Set Performance [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>PPV</td>
<td>R</td>
</tr>
<tr>
<td>1</td>
<td>62.3</td>
<td>89.0</td>
</tr>
<tr>
<td>2</td>
<td>61.8</td>
<td>73.2</td>
</tr>
<tr>
<td>3</td>
<td>70.7</td>
<td>92.4</td>
</tr>
<tr>
<td>4</td>
<td>62.8</td>
<td>87.2</td>
</tr>
<tr>
<td>5</td>
<td>70.9</td>
<td>90.1</td>
</tr>
<tr>
<td>6</td>
<td>70.8</td>
<td>91.5</td>
</tr>
<tr>
<td>7</td>
<td>71.1</td>
<td>90.2</td>
</tr>
<tr>
<td>Baseline</td>
<td>62.0</td>
<td>61.4</td>
</tr>
</tbody>
</table>

✓ variable set used; × otherwise. Bold numbers indicate the best results. The p value was calculated by comparing the AUC between the best algorithm (Set 7) and the other algorithms using paired t-test. N/A indicates that the performances between the two algorithms are identical and no p value is returned.

Table 3. Variables output by logistic regression that were significant at p<0.1 level (ordered by odds ratio).

<table>
<thead>
<tr>
<th>Variable description</th>
<th>OR 95% CI</th>
<th>Variable description</th>
<th>OR 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease specific trial: yes vs no*</td>
<td>17.5 2.74-111</td>
<td>PCT of house at &lt;50% pov. line*</td>
<td>0.84 0.70-1.00</td>
</tr>
<tr>
<td>Type: intervention. vs observation.*</td>
<td>2.38 1.12-5.09</td>
<td>Arrival means: by car vs walk in*</td>
<td>0.78 0.61-1.00</td>
</tr>
<tr>
<td>Arrival season 1: winter vs summer*</td>
<td>1.49 1.12-1.97</td>
<td>Multi-center trial: yes vs no</td>
<td>0.48 0.20-1.15</td>
</tr>
<tr>
<td>Disposition: discharge vs admission*</td>
<td>1.37 1.04-1.80</td>
<td>Follow up call: yes vs no*</td>
<td>0.12 0.03-0.50</td>
</tr>
<tr>
<td>Arrival season 2: autumn vs summer*</td>
<td>1.30 0.96-1.75</td>
<td>Invasiveness 1: Lv.5 vs Lv.1*</td>
<td>0.11 0.02-0.66</td>
</tr>
<tr>
<td>Race: White vs African American*</td>
<td>1.26 1.01-1.59</td>
<td>Invasiveness 2: Lv.2 vs Lv.1*</td>
<td>0.10 0.02-0.49</td>
</tr>
<tr>
<td>Length of stay*</td>
<td>1.03 1.00-1.06</td>
<td>Follow up visit: yes vs no*</td>
<td>0.05 0.00-0.39</td>
</tr>
</tbody>
</table>

*variable significant at p<0.05 level; + variable significant at p<0.1 level. OR: odds ratio; CI: confidence interval.

References

Comparing Weight Redistribution and Distance Imputation Methods for Missing Data in Clear-text and Encrypted Record Linkage
Toan C. Ong, PhD1, Lisa M. Schilling, MD, MSPH2, Michael G. Kahn, MD, PhD1
Departments of Pediatrics1 and Medicine2, School of Medicine
University of Colorado Denver, Colorado USA

Introduction
Complete data are necessary to perform high-quality comparative effectiveness and outcomes research. Complete datasets can be created without sharing clear-text identifiers by performing privacy-preserving record linkage (PPRL) on distributed datasets that capture data from different sources (e.g. clinical, claims) for overlapping patients. [1-4] PPRL is usually performed by comparing one or multiple common encrypted linkage variables, also called quasi-identifiers.[5] Probabilistic record linkage methods assigns a weight to each linkage variable, which captures the ability of that variable to discriminate correct from incorrect linkages.[6] The incorporation of distance measurements, such as dice coefficient, allows PPRL methods to be tolerant of data incorrectness.[7] A similarity score representing the similarity between two records is computed based on the weight and distance between linkage variables. A match is determined by comparing the similarity score to a threshold, called the acceptance level. Despite being able to deal with incorrect data, current clear-text and encrypted record linkage methods usually do not have a robust mechanism to deal with missing linkage data. For example, the open-source FRIL system offers the FRIL-0 option that assigns zero distance to linkage variables with missing data, which essentially declares all missing variables to be non-matching.[8] In a recent publication, we demonstrated the superior performance of two novel methods, Weight Redistribution (WR) and Distance Imputation (DI), to solve the problem of missing linkage data in clear-text record linkage.[9] We applied the same methods to both clear-text and encrypted data and compare their performances with the FRIL-0 method. Since encrypted data require different methods to measure the distance between two values, the effectiveness of the previous methods is not guaranteed.

Methods
Four datasets, which were divided into 2 groups, were used to test the WR and DI methods. Group 1 contained 2 datasets with clear-text data and group 2 contained 2 datasets with encrypted data. Each dataset contained 5000 records and pairs were designed to have exactly 3000 match records. The clear-text data were corrupted to simulate typographical errors (e.g. letter deletion, insertion, and transposition) and values were randomly removed entirely to simulate missingness. The average corruption rate and missing rate in each field was approximately 10%. To ensure comparability, the corrupted clear-text data were used to create the encrypted dataset using two encryption steps: 1) hashing clear-text value with a single 256 character random string salt and 2) adding hash values to a Bloom filter. [10] The linkage fields used for all methods include first name, last name, birthdate, sex, and city of birth. Sex was not used as a linkage variable in our previous publication. Refer to http://goo.gl/ODgMZb for more detailed information about the WR and DI methods and examples with clear-text data.

Results
We compare the performance of three methods with different acceptance levels to deal with missing linkage data in both clear-text linkage and PPRL (Table 1). WR and DI had a much higher overall number of true positive (TP) than did FRIL-0. All methods have positive predicted values (PPV) greater than 0.96 but they have marked differences in sensitivity with both WR and DI having greater sensitivity than FRIL-0. Using the methods on encrypted data sets, derived from the clear-text data sets, did not decrease the sensitivity of the methods. With both clear-text and encrypted data, raising the acceptance threshold eliminated the false positives (FP) but also decreased the sensitivity of the match, resulting in more false negatives. For each dataset type with equal acceptance thresholds, DI appears to have the highest number of TP matches and fewest FP cases.

Discussion
The results showed that the WR and DI methods, which have mechanisms to deal with missing linkage data, outperformed the FRIL-0 method in all performance measures. However, in some cases both of these methods produced FP matches, which would incorrectly merge records from two different individuals. With clear-text data, although the WR method with acceptance level = 80 (row #2) produced 3 FP matches, it was the most effective method with the highest number of TP matches and low FP matches. In addition to increasing the acceptance level to limit the number of FP cases, another solution could be to use more discriminating fields as linkage variables. With encrypted data, including more discriminating linkage variables will also reduce the number of FP cases. The results showed that WR and DI also provide a promising solution for missing linkage data in PPRL.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Acc level</th>
<th># of TP</th>
<th># of FP</th>
<th># of FN</th>
<th>Sensitivity</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear-text</td>
<td>FRIL-0</td>
<td>80</td>
<td>1,594</td>
<td>0</td>
<td>1,406</td>
<td>0.531</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>WR</td>
<td>80</td>
<td>2,751</td>
<td>3</td>
<td>249</td>
<td>0.917</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>WR</td>
<td>100</td>
<td>2,353</td>
<td>0</td>
<td>647</td>
<td>0.784</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>80</td>
<td>2,546</td>
<td>0</td>
<td>454</td>
<td>0.849</td>
<td>1.000</td>
</tr>
<tr>
<td>Encrypted</td>
<td>FRIL-0</td>
<td>80</td>
<td>1,534</td>
<td>0</td>
<td>1,466</td>
<td>0.511</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>WR</td>
<td>80</td>
<td>2,764</td>
<td>113</td>
<td>236</td>
<td>0.921</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>WR</td>
<td>95</td>
<td>2,272</td>
<td>0</td>
<td>728</td>
<td>0.757</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>80</td>
<td>2,445</td>
<td>2</td>
<td>555</td>
<td>0.815</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>DI</td>
<td>90</td>
<td>2,312</td>
<td>0</td>
<td>688</td>
<td>0.771</td>
<td>1.000</td>
</tr>
</tbody>
</table>

TP = true positive; FP = false positive (incorrect matches); FN = false negatives (incorrect non-matches); Sensitivity = TP/(TP+FN); PPV: Positive Predictive Value = TP/(TP+FP);
WR = Weight Redistribution; DI = Distance Imputation; Acc level = Acceptance level.

### Table 1 – Performance of record linkage methods

#### References

#### Abstract

Missing data in record linkage variables can negatively impact the effectiveness of privacy protected record linkage (PPRL). Weight Redistribution and Distance Imputation are novel methods developed to deal with missing data in record linkage variables. Using synthetic datasets with corrupted and missing data we showed that the performance of clear-text and encrypted probabilistic record linkage was improved and not markedly different with encrypted variables. These methods provide promising solutions to missing linkage data in PPRL.
Quantifying Tobacco Exposure Using Clinical Notes and Natural Language Processing to Enable Lung Cancer Screening

Travis J. Osterman DO1, Wei-Qi Wei MMed, PhD1, Joshua C. Denny MD, MS1
1Vanderbilt University, Nashville, TN

Introduction
Lung cancer is the leading cause of cancer-related mortality worldwide with 1.2 million deaths annually1. The National Lung Cancer Screening Trial demonstrated a reduction in lung cancer-associated mortality through annual screening of high risk patients2,3. Based in part on those results, the United States Preventative Services Task Force (USPSTF) recommends annual screening for patients age 55-80 that have 30 or more pack-years (PY) of tobacco use4. Previous research has focused on identifying smokers from EHR data. Here we propose an algorithm to calculate lifetime tobacco exposure in PY using unstructured narrative text from the social history (SHx) section of clinical notes.

Methods
This study used Vanderbilt University Medical Center’s de-identified electronic health record (EHR) with institutional review board approval (IRB #142018). A training set of 250 individuals with a history of smoking was identified from the EHR using a previously validated algorithm5. A validation set of 1000 individuals was selected randomly from the EHR. The SHx from each encounter is stored in the EHR as unstructured free text. The most common, non-blank, social history for each patient was considered the SHx for this study. Sixteen individuals with no available SHx were removed from prior to analysis.

Each individual’s SHx was manually reviewed by two physicians to quantify tobacco usage in PY. This served as the ground truth. We implemented a tiered, rule-based, system using regular expressions (RE). The first rules determine whether the sentence discusses smoking. The second set of rules determines whether PY calculation is possible. A third set of rules extracts rate, duration, and/or total PY. Finally, a PY duration is calculated.

Precision, recall, and F-measure are reported for the second set of rules which identified text with adequate data for extraction (ie: rate and duration, or PY). Tobacco exposure in PY was analyzed using Pearson correlation and root mean square error (RMSE). A one-tailed exact binomial test (α = 0.05) was used to analyze individuals meeting USPSTF screening criteria as previously described and reported using a p-value.

Results
A total of 984 individuals’ SHxs were used for validation (Table 1). There were 162 (16%) smokers in this group. Fifty-three (5.4%) individuals had sufficient information to calculate PY. Mean PY was 31.5 with a range [0.43 – 104]. Nineteen individuals qualify based on USPSTF age and PY criteria. Forty-six of 53 SHxs were correctly classified as having sufficient data to calculate PY (1 false positive, 7 false negatives, precision: 0.98, recall: 0.87, and F-measure: 0.92).

The model made PY predictions on 47 individuals with a Pearson correlation of 0.78, 95% confidence interval [0.65 – 0.87], (Figure 1) and RMSE of 8.1. Fourteen of 19 individuals were successfully identified as lung cancer screening candidates (exact binominal test p-value 0.03).

Discussion
We developed an algorithm to quantify tobacco exposure in PY and report a strong correlation between our prediction and physician review. Using this tool, we were able to identify individuals who qualify for USPSTF lung cancer screening. However, several challenges arose while working with these data. First, these rules are based on a loose vocabulary that includes non-standard abbreviations and misspellings which may be medical center or regionally dependent. Also, our algorithm did not consider when former smokers had quit smoking. Finally, we did not parse the entire clinical note for smoking information outside of the SHx

Conclusion
We have shown the problem of quantifying tobacco exposure can be addressed with common informatics tools. This model can be used to identify potential candidates for lung cancer screening at other medical centers.
Table 1. Population characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>validation set size</td>
<td>1000</td>
</tr>
<tr>
<td>patients with at least one social history entry</td>
<td>984</td>
</tr>
<tr>
<td>Ever-smokers</td>
<td>162</td>
</tr>
<tr>
<td>Tobacco quantify documented</td>
<td>53</td>
</tr>
<tr>
<td>PY range</td>
<td>0.43 - 104</td>
</tr>
<tr>
<td>PY mean (SD)</td>
<td>31.5 (28.0)</td>
</tr>
</tbody>
</table>

Figure 1. Predicted vs physician-calculated tobacco exposure from social history.
* calculated for 14 of 19 successful screening classifications using binomial exact test, one-tail \( \alpha = 0.05 \)
** Pearson correlation 0.78, 95% confidence interval [0.65-0.87]

References
Title: Patient Access to Electronic Health Records During Hospitalization: High Expectations, Fewer Problems, and Fewer Benefits Than Expected

Authors: Jonathan Michael Pell, MD, Assistant Professor of Medicine, Department of Internal Medicine, Division of General Internal Medicine, University of Colorado Denver, School of Medicine
Mary Mancuso, MA, PRA, University of Colorado Hospital, Professional Resources, Aurora, CO, USA
Shelly Limon, BSN, CNRN, Neuroscience Nurse Manager, University of Colorado, Aurora, CO, USA.
Kathy Oman, RN, PhD, CEN, FAEN, Associate Professor, Adjunct, University of Colorado Denver, College of Nursing, Aurora, CO, USA.
Chen-Tan Lin, MD, FACP, Professor of Medicine, Department of Internal Medicine, Division of General Internal Medicine, University of Colorado Denver, School of Medicine, Chief Medical Information Officer, University of Colorado Health, Aurora, CO, USA.

Introduction- In 2001, the Institute of Medicine recommended improving patient engagement by providing care that is continuous, allowing patients to be the source of control, and fostering transparency with patients and families¹. Electronic health records (EHRs) facilitate these objectives via the use of patient portals. Giving outpatients direct access to their health information helps clinicians find errors and improves patient satisfaction, although the implications of this type of access has not been well studied in the inpatient setting². This hospital-based study evaluates the patient, clinician, and nurse experience with immediate (real-time) release of test results and other EHR information through a patient portal.

Methods- This prospective cohort study was performed on a medical unit at a 412-bed, academic tertiary care hospital, from October 2012 to March 2013. Participants were hospital clinicians, nurses, and patients. Patient participants were enrolled by convenience sampling and used a study-provided electronic tablet to access parts of their EHR, including medication schedule, and test results. Patients, clinicians, and nurses completed a survey before and after the intervention. The survey evaluated the domains of caregiver workload, patient confusion and worry, patient empowerment, error finding, and discharge planning. McNemar’s test was performed to analyze binary data between paired responses on surveys for all groups.

Results- Response rates for individuals completing both pre- and post-intervention surveys were 100% for patients (n=50/50), 93.3% for clinicians (n=28/30), and 87.5% for nurses (n=14/16). Patient portal use (defined as clicks/day) was 15.6 (SD=16.2; median=11.2; range=0.26-86.8), and time logged on ranged from 2-1331 minutes. We did not assess the use of the tablet for other purposes or by other users. Figure 1 shows the pre- and post-intervention survey results. The majority (≥65%) of clinicians and nurses were concerned that giving patients immediate access to their test results would increase their workload, but this sentiment decreased in both groups post-intervention. Concerns that seeing test results would cause patient worry were high (>85%) among clinicians and nurses and greater than among patients pre-intervention, but these concerns decreased in all groups. A majority of patients endorsed empowerment items including control, understanding, reassurance, following recommendations, and trust both pre- and post-intervention.
Clinicians (96%) and nurses (93%) were more optimistic than patients (44%) that patient access to their medication lists would help them find errors, and this decreased significantly across all groups post-intervention (patients, -38%; clinicians, -31%; nurses, -43%). Prior to the intervention, most patients (67%) indicated that they would better understand when they would be discharged, but after the intervention, the number of patients endorsing this item fell significantly (-42%, P<.001).

**Discussion** - The suspected risks of giving inpatients direct access to their EHR did not bear out with no increase in workload reported by nurses or clinicians and no increase in confusion or worry reported by patients. Consistent with outpatient studies, patients answered more positively to empowerment questions after being given EHR access. Despite supporting patient empowerment, the promise of patients finding errors in their medications or knowing when they were being discharged never materialized. This is the first published study we are aware of that has evaluated the experience of a large sample of inpatients and their front-line care providers with real-time inpatient EHR access, although it involved patients and providers on a single hospital unit. Federal programs recommend that patients be able to access results from their hospitalization within 36 hours of discharge. This requirement still misses an opportunity for patient engagement through better transparency, and future policies should consider real-time EHR access for inpatients.

**Figure 1.** Pre- and Post-intervention Survey Results for Patients, Clinicians, and Nurses.

**References**
Uncovering the Cognitive Demands of EHR Use via Task Analysis

Mark S. Pfaff, PhD1, Ozgur Eris, PhD1, Amanda Anganes, MS1, Mohammad Rahman, MA1, Tina Crotty, MA1, Jonathan Nebeker, MS MD2, Merry Ward, PhD3

1The MITRE Corporation, Bedford, MA. 2Salt Lake City VA Medical Center, Salt Lake City, UT. 3Department of Veterans Affairs Office of Informatics and Analytics, Raleigh, NC.

Introduction

Using an electronic health record (EHR) system places many cognitive demands upon the user. Complicating an analysis of these demands is the heterogeneity of the many user roles and the vast scope of system functionality. Expert clinicians are highly adaptive to the constraints of the EHR as they develop strategies to seek, share, and store information.1 In this study, 10 clinical experts in a variety of clinical roles were interviewed to elicit the types of challenges that EHR users face and the cues and strategies they employ to overcome those challenges.

Methods

Applied Cognitive Task Analysis (ACTA) was chosen as the method based on its structured approach for eliciting cognitive demands and skills from expert decision makers.2 In particular, ACTA captures expertise in a way that facilitates its transfer to trainees and translation into design recommendations for interactive systems. Interviewees included both inpatient and outpatient practitioners using a variety of EHR platforms: four primary care physicians, two nurse practitioners, one registered nurse, one physician’s assistant, one clinical pharmacist, and one patient care coordinator. The average years of clinical experience was 23.9 (range 8-35). Each interview was jointly conducted by two researchers and lasted approximately two hours.

Results

Using a grounded approach, several themes emerged from the interviews, as presented in Table 1. The reported issues refer to specific EHR implementations and are not necessarily global issues with all EHRs. Nine of out the ten participants reported that they rely 100% on their EHR to perform their clinical tasks. The exception was an ER nurse using a 14 year-old system that served as more of a distraction from other duties than an aid in her work.

Discussion

Fragmentation of information either within or between systems was the top complaint from participants. Requiring users to seek and integrate information from multiple sources requires slow effortful thinking and extensive knowledge of how to navigate the system, whereas displays that present all necessary information in one place would enable rapid associative thinking.3 “What if?” tools were frequently noted as a desired feature, indicating limited support for planning in current EHRs. Ordering medications was a recurrent point of pain for many of the participants. This problem represented the intersection of information overload (vast lists of available medications) with the heightened possibility of entry error (e.g. selecting one above or one below the desired medication) and the needle-in-a-haystack search for a medication that turns out not to be in the system at all.

Communication was also a common sticking point. It is generally hard to know whether someone has received or read important information about the patient – one must rely on trust that other members of the team know what they need to know or have completed expected tasks. This reflects problems in establishing common ground4 and ensuring the team understands the current care goals and the knowledge and responsibilities held by team members.

This research aimed to understand the cognitive demands from a variety of clinicians’ EHR use in different clinical environments and roles. Some of the previously identified cognitive consideration categories for EHR design, such as support for an accurate situation model of the patient, common ground, and planning and action, were encountered and detailed. Ongoing data analysis will further substantiate and structure those dimensions.

This work was funded by the Department of Veterans Affairs (VA) Office of Informatics and Analytics/Health Informatics/Vista Evolution (Projects VA118A-13-D-0037 and VA118A-14-J-0194). Approved for Public Release; case #15-0601.
Table 1. Selected cognitive demands of using an EHR extracted from the ACTA interviews

<table>
<thead>
<tr>
<th>Cognitively demanding issue</th>
<th>Why is it difficult?</th>
<th>Expert strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient information is scattered throughout the same EHR system.</td>
<td>Fragmentation within the EHR results in clinicians engaging in “detective work” in order to construct a holistic and accurate picture of a given patient. This can cause critical pieces of information to be unavailable during care. For instance, lab results associated with an initial acute encounter are not linked to the chronic condition which consequently develops.</td>
<td>Develop an accurate mental model of the topography of the EHR coupled with a sophisticated understanding of the scenarios that invoke the use of its different parts and modules. Clinicians repeatedly stated that “knowing where to look” within the EHR is required to find connections among data.</td>
</tr>
<tr>
<td>The EHR module is not matched to the clinical task, and it functions with inaccurate risk/benefit assumptions.</td>
<td>“One size fits all or provides the most coverage” approach to EHR application can cause a significant mismatch between the clinicians’ needs and the EHR’s functionality. For instance, this can happen when a nurse is forced to fill in several screens in order to be able to order blood while a patient bleeds out during surgery in the OR. The criticality of the information might be justified in a less risky operational environment, but is unjustified in an emergency situation in the OR.</td>
<td>Employ informal communication or task execution mechanisms that do not involve the EHR. Bypass the EHR to the best extent possible. In the OR case, the nurse uses the phone to place the order although she is still required to follow up with the EHR to place the official order (still poor use of critical OR time).</td>
</tr>
<tr>
<td>Effortless generation of effortful downstream work through EHR’s automation capability.</td>
<td>Clinicians operating upstream in the clinical workflow have the ability to create generic care order templates that are specific to a condition, and automatically apply them to multiple patients sharing similar conditions with very little effort. However, not every single order on a template is relevant to a given patient, which results in information overload and time pressure in a pharmacy or laboratory.</td>
<td>If operating downstream in the clinical workflow, take control of existing order templates and the process of creating new templates (e.g. administratively limit clinicians’ ability to overuse automation).</td>
</tr>
<tr>
<td>Asynchronous electronic communication capability of EHRs</td>
<td>Clinicians can send messages to each other through the EHR, but it is difficult for a clinician to read and respond to each message in the presence of other more pressing responsibilities. This results in communication overload.</td>
<td>Try to identify and only respond to important EHR messages, ignore the rest, and revert to informal synchronous communication mechanisms such as the telephone to execute high-priority tasks.</td>
</tr>
<tr>
<td>Medication ordering presents overwhelming choices</td>
<td>In the interest of being comprehensive, pull-down medication lists include many variations of a drug, making selection a possible source of data-entry error.</td>
<td>Develop custom lists of “favorite” or frequently ordered medications.</td>
</tr>
</tbody>
</table>

References

Status of the Health Center-Controlled Network Program: Advancing Health Care Quality through Health Information Technology at Community Health Centers

Anna Poker, RN, MS; Derrick Wyatt, RN, MS; Dominick Black; Suma Nair, MS, RD; Beth Perrine, MA; Tracey Orloff, MA

Author affiliations: AP, DW, DB, SN, BP and TO – Health Resources and Services Administration, Rockville, Maryland

Introduction and Background:

Health Center-Controlled Networks (HCCNs) are consortia of community health centers (HCs) funded by the Health Resources and Services Administration (HRSA), an agency within the U.S. Department of Health and Human Services. In FY2013 HRSA funded a total of 43 HCCNs to advance health care quality through health information technology (health IT). This 3 year HCCN program totals $21 million/year has three main goals: 1) to advance the adoption of ONC certified electronic health records (EHRs,) 2) to promote the participation of eligible providers (EPs) in the EHR incentive program from the Centers for Medicare & Medicaid Services commonly known as “Meaningful Use,” and 3) to improve quality through achieving Healthy People 2020 goals and through recognition of HCs as Patient-Centered Medical Homes (PCMHs).

Methods: We analyzed grant grantee application information to collect baseline data for all the projects. We compiled descriptive data about the number of HCs and the number of EPs across the HCCN program. The baseline data about each HCCN grant project, including the percentage of EPs participating in Meaningful Use, the percentage of HCs that have achieved the Healthy People 2020 goal on at least one measure reported in the Uniform Data System (UDS)¹, and the percentage of HCs that have received PCMH recognition is compared with yearly progress reports.

Results: The HCCN grant program consists of 43 network grantees, which corresponds to 768 HRSA-funded HCs, 3,554 clinical sites, 13,928 clinical providers, and over 12 million patients served. Data collected from Year1 of the Progress Reports reveal that the mean percentage of clinical sites that have adopted EHRs is 90.2 % (from 82%). The mean percentage of EPs who have registered and attested/applied for Meaningful Use payments is 76% (from 66%), while the mean percentage of EPs who have received Meaningful Use payments is 69% (from 56%). The mean percentage of HCs that have achieved the Healthy People 2020 goal on at least one UDS clinical quality measure is 90% (from 75%). The mean percentage of HCs that have achieved PCMH recognition is 42% (from 17%).

Discussion and Conclusion: The HCCN grant program has substantial reach and shows advancements in propelling health centers toward improved quality of health care delivery for millions of underserved and medically vulnerable people served by HRSA-funded HCs. Although HCs are advanced in terms of health IT adoption there is great opportunity for them to optimally use their EHRs and health IT. The HCCNs are providing support for HCs to capture, analyze, and return actionable data for quality improvement at the HC level and to improve population management endeavors.

¹ The Uniform Data System (UDS) is a core set of information appropriate for reviewing the operation and performance of health centers. Health centers report UDS data to HRSA annually.
Interim Results of a Randomized Controlled Trial on Inpatient Engagement

Jennifer E. Prey, MPhil, MS1, Beatriz Ryan, MPH1, Min Qian, PhD2, Ruth M Masterson Creber, PhD, MSc, RN1,3, Susan Restaino, MD4,6, Suzanne Bakken, PhD, RN1,3, Steven Feiner, PhD5, Rebecca Schnall, PhD, RN3, George Hripcsak, MD, MS1,6, Jungmi Han1, David K. Vawdrey, PhD1,6

1Dept. of Biomedical Informatics, 2Dept. of Biostatistics, 3School of Nursing, 4Dept. of Medicine, 5Dept. of Computer Science, Columbia University, New York, NY; 6NewYork-Presbyterian Hospital, New York, NY

Introduction

The impact of patient engagement has been likened to that of a “blockbuster drug.” Higher levels of patient engagement may improve patient satisfaction, patient–provider communication, and adherence to treatment plans. While patient engagement has been a focal point in the outpatient setting, little work on this topic has been conducted within the hospital. This podium presentation will report interim results of a randomized controlled trial (RCT) that evaluates the impact of a tablet-based patient portal tool in the hospital setting.

Methods

Beginning in summer 2014, patients on cardiology wards at Columbia University Medical Center were enrolled in a three-arm RCT, receiving either: 1) usual care, 2) a tablet computer with basic access to the internet, or 3) a tablet computer with access to a personalized patient portal, Inpatient myNYP. Patients responded to a survey consisting of satisfaction and usefulness items, and care team member awareness, along with the Patient Activation Measure (PAM). The PAM classifies individuals into four hierarchical levels, from Level 1 (“Disengaged and overwhelmed”) up to Level 4 (“Maintaining behaviors and pushing further”). Five-point Likert-type scales were used to assess patients’ agreement with survey items (1=“Strongly Disagree” to 5=“Strongly Agree”). Descriptive statistics were calculated for Likert-type survey questions. A within-subjects ANOVA model was used to analyze differences of patients’ perceived ability to identify their care team members. Analyses were performed using R.

Results

Through February 2015, 45 patients have been recruited into the RCT (Table 1). The majority of the participants are male (73.3%), and White (57.8%), with an average age of 63.9 (SD=13.6). At baseline, most participants (51.1%) scored in the third level of the PAM. The distribution of key Likert-type survey question responses can be seen in Figure 1. Results were mostly positive (“Agree” to “Strongly Agree”) to questions about hospital care and care team satisfaction. Figure 2 shows the responses of the perceived knowledge of care team member types. The ANOVA model for the three questions in Figure 2 found an estimated mean score of 4.11, 3.61 and 3.27, respectively (p-value <0.0001, with all three pairwise comparisons also statistically significant), indicating patients were most likely to know the identity of their nurses, followed by their physicians, followed by allied care providers.

Discussion

The positive trend of responses to the survey questions suggests that study participants were generally satisfied with their care in the hospital and believed their care team members listen and encouraged them to participate in their care. However, over 25% of patients chose “Strongly Disagree” or “Disagree” in response to, “I am more involved in my healthcare than I was before my hospital stay.” This result underscores the importance of education and communication during the hospital stay. Knowledge transfer and engagement during this critical period can help patients regain their sense of control before they are discharged home. The results of the ANOVA model, which showed that patients were better able to identify their nurses than their physicians and allied health providers, can be interpreted in at least two ways. First, nurses typically provide more direct patient care than other team members, and thus may establish a more memorable relationship with their patients. Second, in the clinical environment where the study was conducted, nursing practice included manually updating a whiteboard with the name of each patient’s current nurse. Overall, patients seemed to lack familiarity with non-nurse care team members. Improving patient knowledge about the identity of care team members could increase patients’ feelings of control over their care.
Table 1. Characteristics of study participants (N = 45).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean Age ± SD</strong></td>
<td>63.9 ± 13.6</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td>33 (73.30%)</td>
</tr>
<tr>
<td><strong>Hispanic</strong></td>
<td>5 (11.1%)</td>
</tr>
<tr>
<td><strong>Black or African American</strong></td>
<td>11 (24.4%)</td>
</tr>
<tr>
<td><strong>American Indian or Alaska Native</strong></td>
<td>3 (6.7%)</td>
</tr>
<tr>
<td><strong>Asian or Pacific Islander</strong></td>
<td>1 (2.2%)</td>
</tr>
<tr>
<td><strong>Other/Prefer not to answer</strong></td>
<td>4 (8.8%)</td>
</tr>
</tbody>
</table>

**Patient Activation Measure (PAM)**

- Level 1—Disengaged and overwhelmed: 5 (11.1%)
- Level 2—Becoming aware, but still struggling: 10 (22.2%)
- Level 3—Taking action: 23 (51.1%)
- Level 4—Maintaining behaviors and pushing further: 7 (15.6%)

Figure 1. Patient responses, by percentage, to hospital-care and care team satisfaction questions (HC = healthcare).

Figure 2. Patient-reported awareness of care team members by type.

References

What causes pneumonia? Kinds of Knowledge and the Case for Hybrid Representations

Alan L Rector, MD PhD
University of Manchester, Manchester, UK

Introduction

Why is interoperability so difficult? Why have stable standards for EHRs, terminologies, and clinical decision support proved so elusive? One reason is the failure to distinguish fully between different kinds of knowledge and attempts to extend tools built for one kind of knowledge to use for others for which they are inappropriate.

Methods

The paper uses a seemingly simple example of just three families of knowledge representation and data schema formalisms – 1) logic based – SNOMED & OWL – 2) frames, and 3) UML. However, much of what is said applies to RDF, Archetypes, HL7 CDA etc. The result is a set of requirements for hybrid systems and the interfaces among their components. We explore both theoretical issues in the distinctions and the practical technical issues, focusing on the relationship between open-world axiom-oriented representations – e.g. SNOMED’s description logic and OWL – and closed-world template-based representations – e.g. UML Archetypes, frames, etc. The classification of types of knowledge and their relationship to common formalisms is shown in Tables 1 & 2.

No one representation covers everything, even if we limit ourselves to deterministic symbolic representations. For example, OWL/DLs support composite terms, logical definition and classification but not generalizations. Frames, by contrast, support generalizations, but not composite terms, definitions, or classification. UML interfaces more naturally to frames but, in principle at least, specifies data structures rather than knowledge per se.

“Stretching” any one technology to forms of knowledge for which it was not designed causes problems. In particular, attempts to derive UML schemas from OWL or transform OWL to frames usually founder on their different semantics – open-world axiom-based for OWL vs. closed world template-based for frames and UML. These issues can be mitigated, but only by using nonstandard and non-obvious patterns and transformations¹.

Particular problems arise in understanding the difference between OWL/DL expressions and queries on OWL/DL ontologies – e.g. using the new SPARQL 1.1 OWL entailment regime² or the SNOMED Expression Constraint Specification³. Many interfaces between ontologies and other representations depend critically on such queries – for example, “find all of the subcategories of hypertension not classified under diseases of pregnancy”⁴ is very different from the similar DL expression “Find all subclasses of hypertension that necessarily exclude pregnancy.”⁵⁶

Results

The requirements for a representation environment include support for domain knowledge, information schemas, and data. All three forms of domain knowledge are needed – ontologies, generalizations, and specifics. Most practical development also requires language tools to capture what people actually say and avoid misunderstandings – for example even though “subdural hematomas” can be extracranial, if unqualified, “subdural hematoma” almost always indicates intracranial pathology.

There are fundamental logical issues in combining the axiom-based open-world reasoning needed for ontologies/terminologies with the closed-world representation needed to represent generalizations and used by the most rule environments. Hybrid systems are required. For standards development the interfaces among the various parts of the hybrid representation need to be principled and declarative although in places they will necessarily be heuristic to cope with computationally intractable issues. For example, the ICD 11 linkage between the Ontology and Foundation Component⁴ illustrates an interface between ontologies and other representations using queries. In programming environments, more algorithmic approaches such as Hobo⁵ may be appropriate. There are even more serious theoretical issues in incorporating more radically different representations such as Bayesian Networks.

In short: no one representation is sufficient. OWL is not enough; frames are not enough; UML is not enough. It is time for better hybrid solutions.
### Table 1: Kinds of Knowledge

<table>
<thead>
<tr>
<th>Kinds of Knowledge</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology / Terminology</td>
<td>Definitions &amp; necessary conditions for what something is</td>
<td>Pneumonia is an inflammation of the lung parenchyma involving consolidation of aveoli</td>
</tr>
<tr>
<td>Generalizations</td>
<td>Statements that are sometimes true</td>
<td>Pneumonia may be caused by infection</td>
</tr>
<tr>
<td>Practicalities</td>
<td>Statements about good practice</td>
<td>After culture, start patients with pneumonia on broad spectrum antibiotic.</td>
</tr>
<tr>
<td>Specifics</td>
<td>(OWL/DL A-Box) Frames, Logic programming, RDF</td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td>Data Schema</td>
<td>The rules for storing data about a class of specifics</td>
</tr>
<tr>
<td></td>
<td>Data</td>
<td>The diagnosis shall be stored in a field “diagnosis” with a value taken from the SNOMED CT finding set</td>
</tr>
<tr>
<td>Language</td>
<td>Words &amp; grammar</td>
<td>The words, synonyms and any grammatical rules for expressing or recognising information in natural language.</td>
</tr>
</tbody>
</table>

### Table 2: Summary of interaction between kinds of knowledge and technologies.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Axioms</th>
<th>Templates</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology / Terminology</td>
<td>OWL, SNOMED CT formalism, DLs/OWL… Conceptual Graphs…</td>
<td>Frames (in part)</td>
<td></td>
</tr>
<tr>
<td>Generalizations</td>
<td>Frames, Conceptual Graphs, (RDF, UML)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practicalities</td>
<td>Frames, Logic programming, RDF</td>
<td></td>
<td>Logic programming, Rule ML, Business rules</td>
</tr>
<tr>
<td>Specific (facts)</td>
<td>(OWL/DL A-Box)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td>UML, RDFS, Archetypes, HL7 v3 (Frames)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>RDMS, RDF, SKOS.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>Lexicons, grammars, Software tools, e.g. Gate, UMLS language tools…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### References

Understanding Challenges and Opportunities in Precision Medicine and Interoperability Using Informatics Approaches

Jay Geronimo Ronquillo, MD, MPH, MMSc\textsuperscript{1,2}, Chunhua Weng, PhD\textsuperscript{3}, William T. Lester, MD, MS\textsuperscript{4,5}

\textsuperscript{1}Office of the National Coordinator for Health Information Technology, Washington, DC; \textsuperscript{2}Grinformatics, Haymarket, VA; \textsuperscript{3}Columbia University, New York, NY; \textsuperscript{4}Harvard Medical School, Boston, MA; \textsuperscript{5}Massachusetts General Hospital, Boston, MA

Introduction

The Office of the National Coordinator for Health Information Technology (ONC) promotes initiatives that establish and implement standards for certified health information technology (IT) capable of effectively exchanging diverse health data\textsuperscript{1}. Precision medicine is broadly defined as the application of patient-specific genomic information for highly targeted and effective methods of clinical diagnosis, management, and treatment\textsuperscript{2}. Precision medicine has the potential to leverage health IT in ways that could dramatically improve public and population health, bringing practical genetic information exchange into sharp focus\textsuperscript{3}. We can gain insight into the challenges and opportunities intersecting precision medicine and interoperability by analyzing population-level genetic test data, public nomenclature databases, and accepted standards for exchanging genomic information. The purpose of this study was to 1) build an informatics pipeline capable of integrating diverse datasets describing genetic test information, and 2) gain insight into the current state of precision medicine as it relates to data standardization and exchange.

Methods

The NIH Genetic Testing Registry (GTR) provides a comprehensive description of registered genetic tests offered by various laboratories and organizations for clinical applications\textsuperscript{4}. As a starting point for assessing health IT conformance to standards for encoding genetic test data, we chose the HL7 Version 3 Implementation Guide for Family History Interoperability. The minimal core dataset in this implementation guide requires descriptions from both the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) database as well as the National Center for Biotechnology Information (NCBI). We created an informatics data pipeline in python that identified all clinical genetic test information from the NIH GTR, mapped genetic test data to standard identifiers in the HGNC database and NCBI, and created an integrated dataset for analysis\textsuperscript{5}. Analysis was performed using open source statistical software PSPP version 0.8.4 and Microsoft Excel version 14.2.0 (Redmond, WA).

Results

There were a total of 25472 genetic tests submitted by 240 different organizations to the NIH Genetic Testing Registry, testing for approximately 3632 distinct genes. Of these tests, 23999 (94.2\%) were submitted specifically to the NIH GTR, while 1473 (6.1\%) carried over from the earlier GeneTests Laboratory Directory. Of the tests in the NIH Genetic Testing Registry, 23829 (99.3\%) were focused on germline mutations, 113 (0.5\%) on somatic mutations, and 57 (0.2\%) did not provide this information. All 3632 genes were matched to an approved symbol and identifier from HGNC, and 3594 (99.0\%) of these genes were successfully mapped to NCBI RefSeq identifiers. There were a total of 5319 conditions associated with these tests, with 1136 (21.4\%) currently assigned a SNOMED-CT code. The reported purposes for each genetic test were as follows: 23274 (97.0\%) diagnosis, 147 (0.6\%) drug response, 384 (1.6\%) monitoring, 11119 (46.3\%) mutation confirmation, 151 (0.6\%) pre-implantation diagnosis, 5354 (22.3\%) pre-symptomatic, 609 (2.5\%) predictive, 146 (0.6\%) prognostic, 29 (0.1\%) recurrence, 5668 (23.6\%) risk assessment, 4741 (19.8\%) screening, and 154 (0.6\%) therapeutic management.

Discussion

There were a large number of registered genetic tests for a diverse set of genes focused on diagnosis, mutation confirmation, and/or risk assessment of germline mutations that could be passed to offspring. The majority of genes were successfully mapped to HGNC-approved gene symbols and NCBI RefSeq identifiers, consistent with one fundamental implementation guide’s recommendations for exchanging genetic information in health IT. However, most tests submitted to the registry did not include any assigned SNOMED CT code, highlighting the need for more effective (and required) linking of genetic tests to standard clinical codes that explain medical motivations behind test ordering\textsuperscript{6,7}. As precision medicine evolves from assessing genetic tests to applying sequenced genomes, robust informatics tools can provide valuable insight into the wealth of diverse data describing all aspects of healthcare IT\textsuperscript{8}.
References


An Empirical Analysis of Chaplain Charting Practices to Inform Electronic Health Record Template Redesign
Matthew Sakamoto, BS1, Rebecca Johnson, PhD1, M. Jeanne Wirpsa, MA, BCC1, George Handzo, BCC, CSSBB2, Linda Emanuel, MD, PhD1, Abel Kho, MD, MS1
1Northwestern University, Chicago, IL; 2HealthCare Chaplaincy Network, New York, NY

Introduction
With multiple care providers documenting in the electronic health record (EHR), it is important that these distinct disciplines chart in a way that facilitates intra- and inter-disciplinary communication. The healthcare chaplaincy is integral to understanding spiritual and religious beliefs as part of a comprehensive patient assessment, and to address related matters such as providing emotional support or initiating goals of care discussions1. Currently there are no standards for healthcare chaplaincy documentation. This study aims to describe the current state of healthcare chaplaincy documentation and identify important areas for improvement.

Methods
We conducted this study at an academic tertiary medical center with a Pastoral Care department and chaplaincy training program, where trainees also chart on patients. We looked at the NeuroSpine ICU (NSICU) that has chaplains integrated into clinical rounds. Chaplains use a Powerchart (Cerner) template with structured elements and an unstructured narrative note. We extracted Pastoral Care notes using a text-string search from our Enterprise Data Warehouse (EDW), and restricted our analysis to only experienced healthcare chaplains. We performed a descriptive analysis of healthcare chaplain utilization of structured data fields in the Pastoral Care EHR template. A manual review for themes in free text fields for 100 notes were compared to frequencies of corresponding structured data fields. We also conducted an informal mini-survey of NSICU providers to understand multidisciplinary team interaction with chaplains’ documentation.

Results
Our final analysis included 491 patient notes. Structured data items from the Pastoral Care template are grouped into four core domains and the frequency they were utilized were: Plan of Care, Referral source, Spiritual/Religious Affiliation, Spiritual Assessment. In the structured fields, analysis of the sub categories showed that the most frequently checked categories were: “Address anxiety through pastoral presence” (69%) for Plan of Care, Support systems in SR Resources (92%), Prayer in SR practices (75%), and Anxiety/Fear for SR concerns (81%). Initial thematic coding of narrative notes recorded by chaplains for the same patients suggests both convergence and divergence of topics between the structured and unstructured fields. Family support and specific SR practices were consistent in both structured and unstructured fields. Coders identified additional themes including records of care provided beyond prayer, and notes about the impact of SR care. These additional themes in the free text narrative do not have a corresponding structured data field.

Conclusions
Analysis of template documentation by chaplains in one hospital unit that integrates chaplaincy care identified the most commonly used categories within a spiritual and religious Plan of Care. Comparison of structured template data and the free text narrative notes suggests a richer narrative assessment of patients’ needs and identified several areas of improvement in our current template. We identified some potential key concerns documented by the chaplaincy which may be critical to the complete care of the patient (e.g. Do Not Resuscitate status). This study offers insight into the services provided by chaplaincy at one site, and an informal survey with NSICU providers suggest that the chaplain note is valued by the interdisciplinary team for eliciting spiritual concerns and aiding in goals of care discussions. Our findings have helped shape ongoing efforts to redesign the chaplaincy EHR template to better facilitate communication between the chaplains and the healthcare team, specifically establishing two categories of data: one for exclusive use by chaplains for patient continuity and strictly spiritual needs, and one for the rest of the medical team highlighting “Anxiety and Depression,” “Goals of Care,” “Family Coping” or “Religious/cultural beliefs that impact medical care.”

References
<table>
<thead>
<tr>
<th>Referrals</th>
<th>435 89%</th>
<th>Plan of Care</th>
<th>444 90%</th>
<th>Faith Importance</th>
<th>422 86%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community Clergy</td>
<td>1 0%</td>
<td>Address anxiety through pastoral presence</td>
<td>308 69%</td>
<td>Very Important</td>
<td>120 28%</td>
</tr>
<tr>
<td>Crisis</td>
<td>1 0%</td>
<td>Address anxiety through relaxation</td>
<td>5 1%</td>
<td>Moderately Important</td>
<td>182 43%</td>
</tr>
<tr>
<td>DiscernExpert</td>
<td>48 11%</td>
<td>Address communication issues between patient and staff</td>
<td>21 5%</td>
<td>Not Important</td>
<td>15 4%</td>
</tr>
<tr>
<td>Interdisciplinary Rounds</td>
<td>70 16%</td>
<td>Facilitate expression and management of feelings</td>
<td>141 32%</td>
<td>Not Assessed</td>
<td>105 25%</td>
</tr>
<tr>
<td>Nurse</td>
<td>52 12%</td>
<td>Facilitate Sacraments within 24 hours</td>
<td>127 29%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PatientFamily</td>
<td>215 49%</td>
<td>Provide bereavement support for anticipatory grief</td>
<td>21 5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician</td>
<td>8 2%</td>
<td>Provide information and counseling on Advance directives</td>
<td>117 26%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Worker</td>
<td>8 2%</td>
<td>Referral to local faith community</td>
<td>2 0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>83 19%</td>
<td>Provide spiritual counsel for end of life concerns</td>
<td>18 4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>End of Life</td>
<td>34 8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td>70 16%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spiritual Resources</th>
<th>278 57%</th>
<th>Spiritual Practices</th>
<th>156 32%</th>
<th>Spiritual Concerns</th>
<th>228 46%</th>
<th>Anointed</th>
<th>491 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept Illness</td>
<td>26 9%</td>
<td>Meditation</td>
<td>7 4%</td>
<td>Anxiety/Fear</td>
<td>185 81%</td>
<td>Yes</td>
<td>62 13%</td>
</tr>
<tr>
<td>Divine Presence</td>
<td>50 18%</td>
<td>Music</td>
<td>9 6%</td>
<td>Despair</td>
<td>6 3%</td>
<td>No</td>
<td>11 2%</td>
</tr>
<tr>
<td>Hope</td>
<td>65 23%</td>
<td>Prayer</td>
<td>117 75%</td>
<td>Family Conflict</td>
<td>8 4%</td>
<td>N/A</td>
<td>418 85%</td>
</tr>
<tr>
<td>Spiritual WellBeing</td>
<td>54 19%</td>
<td>Ritual</td>
<td>58 37%</td>
<td>Grief</td>
<td>47 21%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support Systems</td>
<td>255 92%</td>
<td>Scripture</td>
<td>20 13%</td>
<td>Loss of Control</td>
<td>83 36%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Need Forgiveness</td>
<td>3 1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spiritual Distress</td>
<td>3 1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unfinished Business</td>
<td>2 1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>End Of Life</td>
<td>28 12%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td>4 2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.** Frequency of Utilization of Structured Data Fields in the Pastoral Care EHR Template
Improvement of cytokine annotation using ontology synonym mapping

Demetrios A. Sarantis, PhD1, Steven H. Kleinstein, PhD1,2, Kei-Hoi Cheung, PhD2,3,
1Department of Pathology, Yale University School of Medicine, New Haven, Connecticut;  
2Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut; 3 Department of Emergency Medicine, Yale School of Medicine, New Haven, Connecticut;

Introduction

The quantity and variety of public data sets in the biomedical domain is rapidly growing, and there is increasing interest in leveraging this prior knowledge through integrative analysis. However, it remains a challenge to discover and integrate such data sets because they are annotated using different terminologies [1]. Mapping annotations to a standardized ontology can help unify terminologies to facilitate data integration. However, single ontologies may have limited coverage of term synonyms and thus can fail to recognize connections between data sets [2]. This research proposes Synonym Ontology Mapping (SOMA), an approach that integrates multiple ontologies to increase synonym coverage [3]. The approach has been tested and validated in the context of annotating cytokines in the immunology literature.

Methods

SOMA maps hierarchically related concepts between ontologies to a common root concept. To demonstrate the approach, we identify ontologies in BioPortal, which contain hierarchies rooted at the cytokine concept. For each of these ontologies, we retrieve all descending classes of the root class utilizing BioPortal search web services. For each descending class, SOMA extracts its identifier, preferred name, and synonyms (Figure 1). We output the extracted identifiers and terms for preferred names and synonyms to a file in RDF format that is a semantic web standard (http://www.w3.org/RDF). The advantage of using RDF graph representation is that we can use RDF tools, such as SPARQL, to query the data and connect them to linked data cloud (http://lod-cloud.net). The graph data are also stored in tabular format in a relational database in order to efficiently implement a synonym-chaining algorithm.

Results

We applied SOMA to the problem of named entity recognition, and constructed the Cytokine Ontology (CYTO) that contains 319 cytokines and 3566 synonyms integrated from ten ontologies available in BioPortal. The testing and validation of CYTO was made by annotating a large number of immunology-related abstracts from PUBMED. Whereas the best single ontology (Gene Ontology) could recognize 169 cytokines in a corpus of >371,000 articles, CYTO recognized 265 cytokines (an 58% improvement) (Figure 2). 727 synonyms, of the above 169 cytokines, would have been recognized using only Gene Ontology, while 1069 synonyms have been recognized using CYTO (an 47% improvement). While there is not any gold standard cytokine ontology, these results demonstrate that the annotation based on SOMA has more coverage than that based on any single ontology.

Discussion

Our work shows that ontologies integration based on synonym mapping increases cytokine term coverage and has the potential to guide development of community standards by identifying frequently used cytokine terms in the literature. During the mapping process (synonym chaining), imprecise synonyms have been identified as a key challenge. Synonym curation is needed to address this issue. Future work will incorporate more cytokine sources (e.g., existing cytokine database/lists) and extend SOMA to additional types of data (e.g., cell populations and cell surface markers). CYTO is publicly available through Bioportal (http://bioportal.bioontology.org/ontologies/CYTO).

References


Figure 1. Graphical representation of cytokine terms and synonyms after synonym integration. Rectangles represent synonyms and ellipses represent terms. Each synonym is connected with a line to its relevant term. Each color depicts terms and synonyms of a specific ontology. Cytokine terms and synonyms that are common to more than one ontology are illustrated with light gray color. (A) All cytokine terms and their synonyms (B) Part of all cytokine terms and their synonyms.

Figure 2. Integrating multiple ontologies increases synonym coverage. (Solid line) The cumulative number of synonyms represented in each ontology. (Dashed line) The cumulative number of synonyms annotated in the literature for each ontology.
A Framework for Individualized Prognosis of Disease Trajectories in Complex, Chronic Diseases: Application to Scleroderma, an Autoimmune Disease

Peter Schulam¹, Colin Ligon², Robert Wise², Laura Hummers², Fredrick Wigley², Suchi Saria¹,³
¹ Department of Computer Science ² Department of Medicine ³ Departments of Biostatistics, and Health Policy and Management
Johns Hopkins University, Baltimore, MD 21202 USA

Introduction: Precision medicine aims to prevent and treat diseases by tailoring management using individual-specific characteristics. Such approaches are especially poised to make an impact in diseases with heterogeneous presentation where treatment plans must vary based on the individual’s disease course. For example, patients with scleroderma, a systemic autoimmune disorder, show a variety of different patterns of disease activity. Individuals can develop complications of the lungs, skin, GI tract, heart, and kidneys to varying extents and at varying rates of progression [1]. Therefore, clinicians must guess which organs should be targeted with aggressive therapy (with potentially harmful side-effects). In many diseases, few biomarkers currently exist to accurately predict an individual’s course. Thus, we propose a computational framework that integrates diverse (static and time-varying) markers to predict the trajectory of a patient’s scleroderma-related lung disease. In addition to tailoring predictions, this dynamic inference procedure implements the vision of a learning health system that adapts and adjusts itself over time.

Methods: We develop a probabilistic framework that represents the disease activity trajectories at multiple resolutions: the population model captures effects that are shared across all individuals in the population, the subpopulation model captures effects specific to a subtype [1,2] --- a group of individuals with similar disease presentation --- and the individual model captures effects specific to a given individual. Population and subpopulation model parameters are learned offline while the individual-specific parameter estimates are refined as more data about the individual are observed. Personalized predictions are computed in real-time by using all of the data in the individual’s clinical history and marginalizing over the individual-specific parameters. Lung health is measured using a clinical marker called PFVC (% of predicted forced vital capacity). Our model uses the PFVC history of the individual to dynamically update predictions. In addition, it uses individual characteristics including demographic (gender and race) and serologic measurements (ACA and Scl-70 antibody positivity).

Results: Figure 1 shows the predictions obtained using the proposed framework for two individuals. Initial PFVC levels are comparable across both patients. After one year of followup, indicated by points (A) and (B), our model is able to correctly predict that the PFVC of the individual in the first row will remain stable, while the other will experience PFVC decline. After 4 years of data the confidence in each prediction is strengthened in spite of the sharp consecutive drop that both individuals exhibit--indicated by points (C) and (D). In particular, after a transient decrease in PFVC in the top patient at point (C) resulting from an episode of cholecystitis, the predicted PFVC trajectory appropriately continued to predict a largely stable course of lung disease. Using 10-fold cross validation on 672 patients, our model achieves mean absolute errors of 10.37, 8.95, and 6.98 when predicting PFVC values between 8-12 years of followup using 1, 2, and 4 years of data respectively.

Discussion: We have introduced a principled framework for integrating diverse clinical marker data to provide personalized prognosis of an individual’s disease trajectory. The proposed framework leverages the idea of refining predictions by modeling deviations at multiple resolutions --- deviations from the population that are common to a subpopulation and deviations across individuals within a subpopulation.

References:
Figure 1: Top row shows white male, around 40 years old, Scl-70 pos, ACA neg, diffuse skin. Bottom row shows African American male, around 50 years old, Scl-70 pos, ACA neg, diffuse skin. Black dots indicate observed points at time of prediction, red dots indicate actual future points. The blue area represents the most likely predicted trajectory. The light green area represents the second most likely trajectory. Pr(*) = confidence in trajectory expressed as a probability.
Are Meaningful Use Requirements Really Meaningful for Medication Use? Experiences from the Field and Future Opportunities

Sarah P. Slight MPharm PhD PGDip,1,2 Eta S. Berner EdD,3 Bill Galanter MS MD PhD,4 Stan Huff MD,5,6 Bruce L. Lambert PhD,7 Carole Lannon MD MPH,8 Christoph U. Lehmann, MD, FAAP, FACMI,9 Brian J. McCourt BA,10 Michael McNamara MD,11 Nir Menachemi PhD, MPH,12 Thomas H. Payne MD,13 S. Andrew Spooner MD MS FAAP,14 Gordon D. Schiff MD,15,16 Tracy Y. Wang MD MHS MSc,10 Ayse Akincigil PhD,16 Stephen Crystal PhD,17 Stephen P. Fortmann MD,18 Meredith L. Vandermeer MPH,18 David W. Bates MD, MSc.2,15,19


Abstract: The Agency for Healthcare Research and Quality (AHRQ)-sponsored Centers for Education and Research in Therapeutics (CERTs) critically examined the impact of MU policy relating to the use of medications. Stakeholders initially met in June 2014 to discuss the specific issues and developed recommendations to help inform future HIT policy. The consensus was that the MU objectives should acknowledge the diversity of healthcare systems and consider in particular EHR functionalities critical for the accurate prescribing of medications in children.

Introduction: To accelerate the deployment of health information technology (HIT), federal legislation provided substantial incentive payments to clinicians and hospitals for adopting and meaningfully using HIT. Meaningful use (MU) requirements aim to assure that electronic health records (EHRs) are utilized with meaningful effect to improve quality, safety, and efficiency of health care. The Agency for Healthcare Research and Quality (AHRQ)-sponsored Centers for Education and Research in Therapeutics (CERTs) critically examined the impact of the MU policy relating to the use of medications, and have jointly developed recommendations to help inform future HIT policy.

Methods: We gathered perspectives from a wide range of professionals representing academicians, practitioners, policy makers, and senior management officials from different healthcare organizations including and beyond the CERTs. Specific issues and recommendations were presented, discussed, and agreed upon as a group. We used the principles of consensus decision-making; all stakeholders were involved in the group discussions, encouraged to contribute opinions and suggestions, given the opportunity to build upon each other’s suggestions, afforded equal input into the process, and allowed to voice any particular concerns that they may have, so that the group could incorporate them into the emerging domains.

Results: Stakeholders’ knowledge and experiences of implementing MU requirements fell into six main domains: accurate medication lists and medication reconciliation, problem list accuracy and the shift in HIT priorities, accurate allergy lists and standards development, support of safer and effective prescribing for children, considerations for rural communities, and general issues with achieving MU. Stakeholders expressed concern about the lack of attention paid to pediatric prescribing, and highlighted the need for standards to better facilitate the exchange of data elements such as patient allergies between healthcare settings. Several organizations felt that their pre-occupation to fulfill MU
requirements stifled innovation at their site. Clinical analysts, engineers and senior programming staff were redirected to work on achieving MU and thus devoted less time to other more meaningful EHR development projects. The structural advantages of existing integrated delivery systems support ready adoption of some Stage 2 recommendations, such as the generation and electronic transmission of discharge prescriptions. However, other measures, such as Summary of Care documentation at time of transitions, required substantial resources to develop significant technical and operational change that affected an exceedingly small fraction of the patient population. Future MU requirements should put more emphasis on flexibly understanding, incorporating, and supporting local health IT configuration that address population health needs.

Discussion: While MU has stimulated EHR adoption, its effects on quality and safety remain uncertain. Stakeholders felt that MU requirements should be more flexible, and recognize that integrated models may be achieving information-sharing goals in alternate ways. Care should be taken to avoid MU requirements that are perhaps unnecessarily burdensome, to the mature, typically staff-model systems that have historically been the leaders in integrated use of clinical information. Better tools and interoperability with external data are needed for effective and efficient medication reconciliation. Future certification rules and requirements should in particular address EHR functionalities critical for the accurate prescribing of medications in children.
Detection of Colorectal Surgical Site Infections Using Bayesian Network and Natural Language Processing

Sunghwan Sohn, PhD1,2, Majid Rastegar-Mojarad, MS1,2,6, James M. Naessens, ScD1,3, Elizabeth B. Habermann, PhD1,3, David W. Larson, MD1,5, Hongfang Liu, PhD1,2

1Department of Health Sciences Research, 2Biomedical Statistics and Informatics, 3Health Care Policy and Research, 4Department of Surgery, 5Division of Colorectal Surgery, Mayo Clinic, Rochester, MN; 6Department of Health Informatics and Administration, UW-Milwaukee, Milwaukee, WI

Introduction
Colorectal surgery (CRS) is performed for various diseases and often requires major reconstruction of the gastrointestinal tract. Because of the number of indications for the surgery and disparities in the level of involvement in colorectal resections, the range of complications differs1. Surgical site infection (SSI) is the most common healthcare-associated infection, accounting for 31% of all hospital inpatient infections2 and resulting in increased length of hospital stay and cost. Over the past decade, accurate identification and monitoring of surgical complications have been emphasized to improve health care quality and to decrease financial costs. The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) is one such effort. Previous studies have examined patient characteristics identified in NSQIP data to assess SSI risk factors1,3,4. These patient characteristics can be regarded as potential factors in predicting SSI, but they are not conclusive factors in identifying SSI. A thorough review of clinical notes is required to determine actual SSIs and this entails a great deal of time and effort. Natural language processing (NLP) and machine learning (ML) techniques represent promising alternatives to automate this process. In this study, we analyzed SSI risk factors for CRS and further applied a Bayesian network classifier to detect SSIs using variables of either or a combination of risk factors in NSQIP data and keywords extracted from clinical narratives by NLP.

Materials and Methods
We used NSQIP data from CRS performed at Mayo Clinic Methodist hospital from 2010 to 2012 combined with SSI indicative keywords from clinical notes of those cases. This consisted of 751 surgery cases with 67 SSIs (i.e., 36 superficial incisional SSIs, 5 deep incisional SSIs, and 26 organ/space SSIs). A Bayesian network was developed to detect SSIs using the following variables: 1) risk factors in NSQIP data; 2) keywords extracted from clinical narratives using NLP; 3) both 1) and 2); and 4) selected variables. The keywords are medical conditions associated with patients and represented by presence or absence. We used an open-source NLP pipeline, MedTagger, which was developed by Mayo Clinic (available in SourceForge), to extract keywords. Detailed descriptions of the variables are found in Table 1.

Results
Table 2 shows the ROC area under the curve (AUC) of the Bayesian network classifier using different sets of variables with 10-fold cross validation. The best performance was achieved when using both risk factors and keywords. The four selected variables, which were based on Chi-squared ranking, were able to produce higher AUC than using only risk factors. Without using these four variables, AUC was 0.729, which is close to using risk factors only. For comparison, a logistic regression was also employed (corresponding ROC curves are shown in Figure 1). The basic statistics and odd ratios of patient characteristics were also examined, and these are described in Table 3.

Discussion
Certain patient characteristics can be used as risk factors to predict SSIs. However, risk factors alone are not sufficient to accurately detect SSIs. The conclusive evidence, such as keywords from clinical narratives, must be examined thoroughly to better identify SSIs. The Bayesian network using both risk factors and keywords produced the strongest performance, although we used a limited set of keywords. Further refinement of keywords and ongoing assessment of risk factors are necessary to better detect SSIs. Automatic extraction of risk factors to implement near real-time SSI surveillance is also a next step. Informatics techniques such as NLP, which allows for the automatic extraction of patients’ medical conditions and statistical ML to predict the corresponding outcomes, are inevitable components in the realization of efficient SSI surveillance.
Table 1. SSI risk factors and indication keywords for colorectal surgery.

<table>
<thead>
<tr>
<th>Potential risk factors</th>
<th>SSI indications in clinical notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre-operation</strong></td>
<td><strong>Intra-operation</strong></td>
</tr>
<tr>
<td>Age; gender</td>
<td>ASA score*</td>
</tr>
<tr>
<td>smoking status</td>
<td>operation type</td>
</tr>
<tr>
<td>pre-op hospital stay</td>
<td>wound classification</td>
</tr>
<tr>
<td>diabetes; anemia†</td>
<td>duration of operation</td>
</tr>
<tr>
<td>BMI (&gt;=30)</td>
<td>laparoscopic vs others</td>
</tr>
<tr>
<td>prior operation</td>
<td></td>
</tr>
<tr>
<td>steroid usage</td>
<td></td>
</tr>
<tr>
<td>blood transfusion</td>
<td></td>
</tr>
<tr>
<td><strong>Concept (keywords):</strong></td>
<td></td>
</tr>
<tr>
<td>Wound infection; Cellulitis</td>
<td></td>
</tr>
<tr>
<td>Abdominal contamination;</td>
<td></td>
</tr>
<tr>
<td>Purulent drainage; Fascial/wound</td>
<td></td>
</tr>
<tr>
<td>Abscess; Infected fluid;</td>
<td></td>
</tr>
<tr>
<td>Antibiotic treatment for infection</td>
<td></td>
</tr>
</tbody>
</table>

†‡ extracted from clinical notes using NLP; † each concept consists of a set of the corresponding keywords

*American Society of Anesthesiologists (ASA) physical status classification

Table 2. Bayesian network AUC for SSI identification.

<table>
<thead>
<tr>
<th>Feature used</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre/intra-op risk factors</td>
<td>0.643</td>
</tr>
<tr>
<td>risk factors</td>
<td>0.721</td>
</tr>
<tr>
<td>keywords</td>
<td>0.799</td>
</tr>
<tr>
<td>risk factors &amp; keywords</td>
<td>0.827</td>
</tr>
<tr>
<td>top four variables†</td>
<td>0.777</td>
</tr>
</tbody>
</table>

†wound infection, abscess, postop antibiotics, purulent drainage

Table 3. Major risk factors for colorectal SSI

<table>
<thead>
<tr>
<th>Patient factor</th>
<th>SSI (N=67)</th>
<th>No SSI (N=684)</th>
<th>Odd Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI (&gt;=30)</td>
<td>37.3% (25)</td>
<td>26% (178)</td>
<td>1.69</td>
</tr>
<tr>
<td>Diabetes</td>
<td>19.4% (13)</td>
<td>8.6% (59)</td>
<td>2.55</td>
</tr>
<tr>
<td>Post-op anemia</td>
<td>40.3% (27)</td>
<td>20.3% (139)</td>
<td>2.65</td>
</tr>
<tr>
<td>Pre-op steroid</td>
<td>16.4% (11)</td>
<td>12.7% (87)</td>
<td>2.55</td>
</tr>
<tr>
<td>Blood transfusion</td>
<td>4.5% (3)</td>
<td>2.8% (19)</td>
<td>1.64</td>
</tr>
</tbody>
</table>

( ) denotes number of cases; % is computed by number in ( ) / N

References

Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP) – a user-centric NLP system

Ergin Soysal, MD, PhD, Jingqi Wang, Min Jiang, Yonghui Wu, PhD, Hua Xu, PhD
School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

Introduction
Natural language processing (NLP) technologies play an important role of unlocking patient information in clinical narratives. Many clinical NLP systems have been developed – e.g. cTakes, MetaMap and MedLEE –, and applied in various applications such as identification of different clinical conditions. Despite the success of these systems, there are still barriers that preventing the widespread use of NLP tools by clinical researchers. For example, applying an existing machine learning module to existing data may not result in desired performances, and specific adaptations like re-annotating and re-training are always needed. However, this is a major obstacle for clinical end users with very limited informatics training. In this paper, we describe CLAMP, a natural language processing toolkit, which is designed to be used by non-technical users for clinical language processing, starting from annotation to training and testing models. The rationale behind CLAMP is to improve experience of end users.

Key Functionality
CLAMP is an NLP toolkit, consisting of an Eclipse based graphical user interface (GUI) (Figure 1) and an open source high performance language processing framework available at https://sbmi.uth.edu/ccb/resources/clamp.htm.

- NLP pipelines: CLAMP components builds on a set of high performance NLP components that were proven in several clinical NLP challenges such as i2b2 (2009-2010, NER #2), ShARe/CLEF (2013, abbreviation #1), and SemEVAL (2014, UMLS encoder, #1). A pipeline can be visually created and customized by chaining CLAMP components in the processing order. Upon creation of a component, CLAMP will check errors and direct user to appropriate logical order for a properly working pipeline. These components are supported by knowledge resources consisting of medical abbreviations, dictionaries, section headers, and a corpus of 400 annotated clinical notes derived from mtsamples.

- Machine learning and hybrid approaches: The framework provides alternative components for some tasks, utilizing rule based methods or machine learning methods using support vector machines, conditional random fields and Brown clustering. These components can be customized by re-training by an annotated corpus, or visually editing the rulesets within the GUI to achieve a custom NLP task. GUI also provides built-in functionality to test the model, using annotated test corpora or n-fold cross validation.

- Corpus management and annotation tool: The user interface accommodates required tools to maintain clinical text corpora. It hosts an improved version of BRAT annotation tool for clinical text annotations.

Evaluation
CLAMP components were tested using a couple of available corpus, suitable for the purpose. Rule based sentence boundary detection, named entity recognition (NER) and assertion components were tested using i2b2 corpus gold standard annotation. Rule based tokenization and section header identification utilizing lexical variant generator with SPECIALIST tested in UT Health corpus against gold standard annotation. OpenNLP based part-of-speech (POS) tagger component tested using 5-fold cross validation with MiPACQ corpus. UMLS encoder component was tested on SemEval-2015 corpus consisting of 431 clinical notes.

Results
Some performance metrics are summarized in Table 1 and Table 2. Accuracy for tokenizer, OpenNLP based POS tagger, section header identification, abbreviation and sentence boundary detection components were found as 0.98, 0.94, 0.96, 0.72 and 0.92 respectively. Overall f-measure for NER component calculated as 0.82 (p: 0.86, r: 0.79). Assertion component showed an f-measure of 0.94 (p: 0.94, r: 0.94. UMLS encoder a performance of 0.73 (accuracy) was reached with a processing speed of 1.5 sec/note.

Discussion
We presented open source CLAMP toolkit that aims to provide the best proven approaches using a mixture of dictionary based, rule based and machine learning methodologies in NLP as an easy to use desktop application for users with minimal or no technical experience. CLAMP toolkit also provides a high performance NLP framework that can be used with UIMA AS to implement more advanced scalable applications.
References

Table 1. Performance of principal CLAMP components

<table>
<thead>
<tr>
<th>Component</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Accuracy</th>
<th>Time per doc (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence Boundary</td>
<td></td>
<td>0.92</td>
<td>0.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tokenizer</td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
<td>4.24</td>
</tr>
<tr>
<td>POS Tagger</td>
<td></td>
<td>0.94</td>
<td>78.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section Header</td>
<td></td>
<td>0.96</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical Abbreviation</td>
<td></td>
<td>0.72</td>
<td>1.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Named entity recognition</td>
<td>0.86</td>
<td>0.79</td>
<td>0.82</td>
<td></td>
<td>1034.43</td>
</tr>
<tr>
<td>Assertion</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td></td>
<td>112.44</td>
</tr>
<tr>
<td>UMLS encoder</td>
<td></td>
<td>0.73</td>
<td>224.77</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Named entity recognition component performances on various semantic classes.

<table>
<thead>
<tr>
<th>Semantic Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>0.85</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>Problem</td>
<td>0.85</td>
<td>0.77</td>
<td>0.81</td>
</tr>
<tr>
<td>Lab tests</td>
<td>0.88</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>Overall</td>
<td>0.86</td>
<td>0.79</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Figure 1. CLAMP provides a complete integrated environment to visually build NLP pipelines including tools for annotation, training and testing models.
PCORnet Implementation of PopMedNet Data Characterization Tools

Jessica L. Sturtevant, MS1, Jessica M. Malenfant, MPH1, Melanie Davies1, Jose Rodrigues1, Laura G. Qualls, MS2, Lesley H. Curtis, PhD2, Jeffrey S. Brown, PhD1

1Harvard Pilgrim Health Care Institute, Boston, MA; 2Duke Clinical Research Institute, Durham, NC

Introduction
The National Patient-Centered Clinical Research Network (PCORnet) distributed research network (DRN) was established by The Patient-Centered Outcomes Research Institute (PCORI) in 2013. The network consists of 11 Clinical Data Research Networks (CDRNs) and 18 Patient-Powered Research Networks (PPRNs). The PopMedNet™ (PMN) platform was implemented across the PCORnet distributed research network in 2014 to help manage and facilitate network operations in support of multi-site clinical trials, comparative effectiveness research and cohort identification and characterization.

The PCORnet Common Data Model (CDM) - developed to facilitate cross-network research - is being adopted by all CRDNs and PPRNs. The PCORnet Coordinating Center (CC) is responsible for data characterization activities related to the CDRN and PPRN implementation of the PCORnet CDM. Data characterization is implemented using a distributed program that is executed locally by each CDRN/PPRN at each local data source. The data characterization output has a standardized structure - a data characterization data model - which allows PCORnet to leverage functionality developed by other distributed networks that use PMN. This data characterization data model is agnostic to the network CDM and allows any number of networks to use the PMN Data Characterization tool. The PMN Data Characterization tool allows for rapid, point-and-click visualization of aggregate metrics across the network.

Methods
Based on the standardized data characterization data model, PCORnet leveraged the PMN Data Characterization tool for race and ethnicity fields that were developed by FDA without any additional software development. Output was obtained from CDRN DataMarts allowing for rapid visualization of the race and ethnicity distribution across the network.

Data characterization metrics that are rapidly accessible will be valuable to the developing PCORnet DRN and will facilitate research. The ability of the PCORnet CC to leverage existing PMN functionality demonstrates the extensibility and cross-network functionality that is a goal of PCORnet. The standardized data characterization data model allows for this cross-network sharing of PMN functionality.

Results
The first PCORnet data characterization queries were designed to capture data elements for race and ethnicity to assess and visualize key characteristics such as missingness and frequency. The PCORnet query was pilot tested to demonstrate the ability to utilize the PMN Data Characterization tool leveraging capabilities originally developed for use by other networks and now embedded in the PMN platform. This pilot demonstrated the ability to quickly implement the PMN Data Characterization tool to visualize the distribution of race and ethnicity among CDRNs.

Conclusion
PCORnet data characterization will encompass over 55 CDRN site data sources and at least 20 PPRN site data sources. Visualizing and describing the characteristics of these diverse data sources that include clinical, claims and patient-reported data is facilitated and simplified in utilization of the PMN Data Characterization tool. This is a novel approach to inform data queries and to investigate data sources in preparation for research requests or clinical trial participant identification.

PCORnet benefits from the cross platform utilization of existing PMN tools and simultaneously contributes new developments that may be leveraged by other networks. Future work will include enhancing the Data Characterization tool to include other relevant data elements, refined access controls and interactive visualization design and reports. Future PCORnet contributions to this functionality will be incorporated into the PMN platform that can then be leveraged by any number of other networks.

Rapid, easy access to high level data descriptive statistics by site and network will help PCORnet assess data usability for varying purposes prior to investing time on code development and data source investigation.
Efficiency and Accuracy of Kinect and Leap Motion Devices Compared to the Mouse for Intraoperative Image Manipulation

Uchenna A. Uchidiuno, MS1, Yuanyuan Feng, MS1, Hamid R. Zahiri, MD2, Ivan George2, Adrian Park, MD, FACS2, Helena Mentis, PhD1

1University of Maryland Baltimore County, Baltimore, MD; 2Anne Arundel Medical Center, Annapolis, MD

Introduction

Touchless interaction devices such as the Microsoft Kinect and Leap Motion device have been proposed for medical image manipulation during surgery1,2. Studies have shown that touchless devices have the potential for alleviating the concerns of maintaining a sterile environment, reducing surgery time, and enabling surgeons to be more autonomous in carrying out various image manipulation tasks required during surgery3. Although these previous studies1-3 have shown the potential of these touchless devices, there is a need to investigate the efficiency and accuracy of these devices compared to the mouse. Prior work has shown that the Kinect can be as efficient and accurate as the mouse, but it is unclear if task type has a mediating effect and these studies have not formally tested the devices with the intended end-user group4,5. In the following study, we carried out a controlled lab-based evaluation of two touchless interaction devices with five different tasks by ten surgeons: the Microsoft Kinect and the Leap Motion. Our study aimed to answer two research questions: 1) how well does touchless interaction compare to a mouse on metrics of accuracy and efficiency and 2) does task type lend itself better to one interaction device over another.

Methods

We used the TedCas plugin for ClearCanvas DICOM viewer to display and manipulate anonymized CT images of a head. TedCas provided the following five image manipulation techniques: step-through, pan, zoom, circle measure, and line measure. The experiment compared three input devices: a standard wireless optical mouse with a sterile covering, the Kinect, and the Leap Motion sensor. Each participant performed each of the five image manipulation techniques five times using each of the input devices. The sequence of tasks and devices were counterbalanced across participants. Data collected was video of each participant captured from the front and a screen capture of the display allowing us to record the completion times and number of errors. A total of ten experienced surgeons participated in the study (see Table 1). The Anne Arundel Health Systems and University of Maryland Baltimore County IRB approved the study and participant consent was acquired. Statistical analysis was performed with SPSS version 22 (Chicago, IL). We carried out a 3 by 2 (Device x Run) ANOVA for both task completion time and errors.

Results

For efficiency, a repeated measure analysis for zoom, pan, line measure and circle measure showed that there is no significant difference between each device in terms of completion time (Figure 1). For the step-through task, there was a significant main effect of Device on Time, F(2, 16)=43.155, p<.001, η2 = .844 (Figure 2). Post-hoc tests using the Bonferroni correction indicated that the mouse was significantly faster than the touchless devices and the Leap Motion was significantly faster than the Kinect. For accuracy, a repeated measure analysis for zoom, step-through, line and circle measure showed that there is no significant difference between each device in terms of errors (Figure 3). The pan task had a significant main effect of Device on Errors, F(2, 18)=4.936, p<.05, η2 = .354. Post-hoc tests using the Bonferroni correction indicated that the mouse had significantly fewer errors than the Kinect or Leap Motion device (Figure 4).

Conclusion

Evaluating the efficiency and accuracy of touchless interaction devices against the mouse gives a better understanding of the benefits and limitations of these devices. We found that the mouse maintains its overall accuracy and efficiency for most tasks. However, for certain image manipulation tasks i.e. zooming, line measure and circle measure, there is no significant difference between the performance of the mouse and the touchless input devices. Both the Microsoft Kinect and Leap Motion device have great potential for adoption in the OR for image manipulation, but there is need for improvement in certain areas if they are to replace the mouse.
Tables and Figures

| Table 1. Participant Demographics. |

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age Range (year)</th>
<th>30-40</th>
<th>40-50</th>
<th>50-60</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Resident</th>
<th>Fellow</th>
<th>Attending</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prior Experience</th>
<th>Leap Motion</th>
<th>Kinect</th>
<th>Gaming</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

**Figure 1.** Average zoom task time by the 10 participants across the 5 attempts with each input device – non-significant difference between the devices.

**Figure 2.** Average stepping through task time by the 10 participants across the 5 attempts with each input device – significant difference between devices.

**Figure 3.** Average number of circle measure task errors for the 10 participants across the 5 attempts with each device – non-significant difference between devices.

**Figure 4.** Average number of pan task errors for the 10 participants across the 5 attempts with each device – significant difference between the devices.

References


5. Juhnke B. Evaluating the Microsoft Kinect compared to the mouse as an effective interaction device for medical imaging manipulations. Grad Theses Diss [Internet]. 2013 Jan 1; Available from: http://lib.dr.iastate.edu/etd/13355.
Physician Participation in Meaningful Use and Rehospitalization of Medicare Fee-for-Service Enrollees

Mark Aaron Unruh, PhD¹, Hye-Young Jung, PhD¹, Joshua R. Vest, PhD¹, Lawrence P. Casalino, MD, PhD¹, and Rainu Kaushal, MD, MPH¹,²,³,⁴, for the HITEC Investigators

¹Department of Healthcare Policy & Research, Weill Cornell Medical College, New York, NY
²Department of Medicine, Weill Cornell Medical College, New York, NY
³Department of Pediatrics, Weill Cornell Medical College, New York, NY
⁴NewYork-Presbyterian Hospital, New York, NY

Abstract: Nearly 20% of hospitalized Medicare fee-for-service enrollees are readmitted within 30 days. Outpatient physicians’ use of interoperable electronic health records may reduce readmission rates. We compared the odds of rehospitalization before and after Meaningful Use for patients of physicians who participated in the program with patients of matched control physicians. Relative to the control group, physician participation in Meaningful Use was associated with 7.5% lower odds of rehospitalization (Odds Ratio: 0.925, 95% confidence interval: 0.860-0.996).

Introduction: Nearly one-fifth of hospitalized Medicare fee-for-service (FFS) enrollees are readmitted within 30 days.[¹] Use of interoperable electronic health records (EHRs) among outpatient physicians may reduce rates of rehospitalization. Meaningful Use has made nearly $30 billion available to encourage provider adoption and use of interoperable EHRs.[²] The objective of this study was to evaluate the impact of outpatient physicians’ participation in Meaningful Use on rehospitalization of Medicare FFS enrollees.

Methods: The study population included 102,857 Medicare FFS enrollees from the State of New York hospitalized over the period 2010-2012. We compared changes in the odds of 30-day rehospitalization before and after the implementation of Meaningful Use for patients attributed to physicians who participated in the program with patients attributed to a matched group of control physicians. Odds ratios were estimated using logistic regression models with hospital and physician fixed effects. Two sensitivity analyses were conducted, one of sicker patients with 3+ Elixhauser co-morbidities and another of patients hospitalized for acute myocardial infarction (AMI), congestive heart failure (CHF), or pneumonia, three conditions targeted by Medicare to penalize hospitals with high rates of readmission.

Results: The characteristics of patients attributed to physicians who participated in Meaningful Use were similar to those of patients attributed to physicians who did not participate in the program during the study period (Table 1). After adjusting for patient, hospital and physician characteristics, patients attributed to physicians participating in Meaningful Use had 7.5% lower odds of rehospitalization (Odds Ratio [OR]: 0.925, 95% confidence interval [CI]: 0.860-0.996) compared to patients attributed to physicians who used paper records. ORs from analyses of patients with 3+ Elixhauser co-morbidities and those with AMI, CHF, or pneumonia were 0.928 (95% CI: 0.869-0.990) and 0.882 (95% CI: 0.771-1.010), respectively. In all analyses, patients of physicians with EHRs who did not participate in Meaningful Use had the same risk of readmission as patients of physicians who used paper records.

Discussion: Physician participation in Meaningful Use may reduce the likelihood of rehospitalization among Medicare FFS enrollees. Patients with conditions associated with high rates of readmission may benefit more from physician participation in the program. Our findings represent the first evidence of the impact of physician participation in Meaningful Use on the quality and utilization of care among the Medicare FFS population.
Table 1. Characteristics of patients, hospitalizations, and physicians

<table>
<thead>
<tr>
<th>Characteristics of attributed patients†</th>
<th>Meaningful Use</th>
<th>Non-Meaningful Use</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total N</td>
<td>53,626</td>
<td>49,231</td>
<td></td>
</tr>
<tr>
<td>Age (mean)</td>
<td>76.0</td>
<td>75.6</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Female (%)</td>
<td>57.1</td>
<td>56.8</td>
<td>0.33</td>
</tr>
<tr>
<td>Race (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>85.6</td>
<td>85.8</td>
<td>0.40</td>
</tr>
<tr>
<td>African-American</td>
<td>9.1</td>
<td>8.7</td>
<td>0.02</td>
</tr>
<tr>
<td>Other race</td>
<td>4.9</td>
<td>5.2</td>
<td>0.08</td>
</tr>
<tr>
<td>Unknown race</td>
<td>0.4</td>
<td>0.4</td>
<td>0.62</td>
</tr>
<tr>
<td>Household income for zip code (mean)</td>
<td>72,351</td>
<td>72,457</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Characteristics of hospitalizations

| Number of Elixhauser co-morbidities (mean) | 2.6 | 2.6 | <.001 |
| Length of stay in days (mean)              | 6.3 | 6.3 | 0.24  |
| Any intensive care unit use (%)            | 14.7| 14.5| 0.31  |

Characteristics of physicians‡

| Total N | 1543 | 1543 |        |
| Female (%) | 23.9 | 25.2 | 0.43   |
| Specialty (%) |      |      |        |
| Primary care | 39.9 | 39.7 | 0.91   |
| Medical      | 36.6 | 36.9 | 0.82   |
| Surgical     | 23.6 | 23.4 | 0.90   |
| Practice size (mean number of physicians)  | 5.0  | 4.9  | 0.79   |

†Characteristics are based on the most recent observation for each patient.
‡Characteristics are based on the most recent observation for each physician.

References

Toward the Development of a Predictive Model for Patient Portal Adoption by Patients in a Federally Qualified Health Center

Theresa L. Walunas, PhD¹, Richard Kalu, MD¹, Matthew Sakumoto¹, Kathryn Jackson, MS¹, Sarah Rittner, MA², Timothy Long, MD²,³, Mita Sanghavi Goel, MD, MPH¹  
¹Northwestern University, Chicago, IL; ²Alliance of Chicago Community Health Services, Chicago, IL; ³Near North Health Services Corporation, Chicago, IL

Introduction
Patient portals provide online access to health records and allow for secure communication with clinicians, test result review, and medication refill requests. Patient portals are critical for achieving meaningful use of electronic health records (EHR) and improving patient centered care, yet studies show disparities in use of patient portals, particularly by African-American populations. In order for health care organizations to improve use of portals, it is important to understand factors that predict adoption and use of portal technologies. The technology acceptance model (TAM) has been used to explain and predict adoption of technology in many environments, but has not been applied to patient portals. To determine if the TAM could be used for portals, we focused on how perceived ease of use (PEOU) and perceived usefulness (PU) of the portal varies by basic demographics and health conditions amongst a patient population at one urban Federally Qualified Health Center (FQHC).

Methods
Prior to implementation of the patient portal at an FQHC with a predominantly African-American patient population, we conducted an in-person survey among a convenience sample of 149 participants. We collected demographic data (age, gender, race/ethnicity, educational attainment, number of existing chronic health conditions in EHR problem list) along with the type and level of technology use. Participants’ ages were categorized as ≤35, 36-50, and 51+ years, and their education as ≤GED, some college, and ≥college graduate. We assessed and compared their PEOU and PU of a portal system. In addition to baseline descriptive analyses, we conducted chi-square tests comparing each domain (PEOU, PU) by demographic characteristics (age, gender, educational attainment, chronic conditions). P<0.05 was considered statistically significant for all analyses.

Results
The mean age of participants was 46 years; 96% were African-American, 61% were women; 28% had a GED or less formal education; 47% had 1 or more chronic conditions. There was a high level of technology use: 78% of participants had a computer, 98% had a cell phone, 89% owned a smart phone, and 83% used the Internet daily. Of the questions measuring PEOU of portal technology, we found few significant differences by age, gender, education or chronic conditions (see Tables 1 and 2). Older participants were more likely to perceive the portal as time consuming (24% of 51+ vs. 5% of ≤35, p<0.05). Men were more likely than women to perceive portals as too complicated to use (30% vs. 14%, p<0.05). Those with 3+ chronic conditions anticipated it would be harder to use compared to those with none (60% vs 87%, p<0.05). Those with higher levels of formal education reported lower rates of PEOU: more reported it was time consuming (25% of college grads vs. 7% of ≤GED, p<0.05) and it was too complicated to use (35% of college grads vs. 7% of ≤GED, p<0.05); fewer reported it was easy to use (73% of college grads vs. 85% of ≤GED, p<0.05) and that learning to operate the patient portal will be easy (68% of college grads vs. 95% of ≤GED, p<0.05). We did not find many differences by age, sex, and education for the PU questions related to features of the patient portal (see Table 3). The few differences were: participants aged ≤ 35 years valued the ability to email their doctors or complete pre-clinic appointment tasks online compared with older participants. Women valued the ability to schedule recommended screening tests, as well as view their clinic notes online, more than men. Finally, patients with 3+ chronic conditions had less value for using the portal to ask questions about medical issues and reviewing current medications compared to patients with fewer chronic conditions.

Conclusions
In this single FQHC site study, we observed an overall high level of technology ownership and use amongst the patients surveyed. PEOU of portal technology varied substantially by level of formal education. Surprisingly, those with higher educational attainment were more apprehensive about ease of use of the system. On the other hand, there were few variations in areas of PU, with most participants placing a high value on communicating with their doctor and managing their care through the portal. Achieving Meaningful Use of EHR and preventing worsening of disparities in portal uptake requires attention to ensuring and emphasizing the ease of use of these systems, especially among those with higher levels of education.
### Table 1: Perceived Ease of Use of Patient Portal Technology by Age and Sex

<table>
<thead>
<tr>
<th>Age (in years)</th>
<th>Sex</th>
<th>Can get help if having difficulty</th>
<th>Too complicated to use</th>
<th>Take too much time</th>
<th>Will be clear and understandable</th>
<th>Not require a lot of mental effort</th>
<th>Easy to use</th>
<th>Easy to have it do my task</th>
<th>Learning to operate will be easy for me</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 35 (n=43)</td>
<td></td>
<td>91%</td>
<td>19%</td>
<td>5%</td>
<td>81%</td>
<td>58%</td>
<td>88%</td>
<td>79%</td>
<td>86%</td>
</tr>
<tr>
<td>35-60 (n=43)</td>
<td></td>
<td>95%</td>
<td>9%</td>
<td>7%</td>
<td>84%</td>
<td>67%</td>
<td>84%</td>
<td>77%</td>
<td>86%</td>
</tr>
<tr>
<td>51+ (n=43)</td>
<td></td>
<td>86%</td>
<td>29%</td>
<td>24%</td>
<td>73%</td>
<td>51%</td>
<td>83%</td>
<td>67%</td>
<td>84%</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td>0.27</td>
<td>0.05</td>
<td>0.01</td>
<td>0.37</td>
<td>0.24</td>
<td>0.71</td>
<td>0.31</td>
<td>0.95</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td>90%</td>
<td>14%</td>
<td>10%</td>
<td>79%</td>
<td>60%</td>
<td>86%</td>
<td>72%</td>
<td>86%</td>
</tr>
<tr>
<td>Men (n=57)</td>
<td></td>
<td>90%</td>
<td>20%</td>
<td>0%</td>
<td>67%</td>
<td>20%</td>
<td>84%</td>
<td>74%</td>
<td>85%</td>
</tr>
<tr>
<td>Women (n=92)</td>
<td></td>
<td>90%</td>
<td>0.57</td>
<td>0.97</td>
<td>0.92</td>
<td>0.79</td>
<td>0.71</td>
<td>0.79</td>
<td>0.84</td>
</tr>
</tbody>
</table>

### Table 2: Perceived Ease of Use of Patient Portals by Education Level and Number of Chronic Health Conditions

<table>
<thead>
<tr>
<th>Education Level</th>
<th>≤ GED (n=41)</th>
<th>Some College (n=68)</th>
<th>College + (n=40)</th>
<th># of Chronic Health Conditions</th>
<th>p-value</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3+</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can get help if having difficulty</td>
<td>95%</td>
<td>90%</td>
<td>85%</td>
<td>0.32</td>
<td>94%</td>
<td>86%</td>
<td>85%</td>
<td>87%</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Too complicated to use</td>
<td>7%</td>
<td>19%</td>
<td>35%</td>
<td>0.01</td>
<td>14%</td>
<td>24%</td>
<td>35%</td>
<td>20%</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Take too much time</td>
<td>7%</td>
<td>10%</td>
<td>25%</td>
<td>0.04</td>
<td>6%</td>
<td>22%</td>
<td>35%</td>
<td>0%</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Will be clear and understandable</td>
<td>85%</td>
<td>81%</td>
<td>68%</td>
<td>0.12</td>
<td>83%</td>
<td>68%</td>
<td>90%</td>
<td>67%</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Not require a lot of mental effort</td>
<td>58%</td>
<td>67%</td>
<td>51%</td>
<td>0.24</td>
<td>57%</td>
<td>51%</td>
<td>75%</td>
<td>53%</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Easy to use</td>
<td>85%</td>
<td>91%</td>
<td>73%</td>
<td>0.71</td>
<td>87%</td>
<td>84%</td>
<td>95%</td>
<td>60%</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Easy to have it do my task</td>
<td>78%</td>
<td>77%</td>
<td>63%</td>
<td>0.21</td>
<td>77%</td>
<td>68%</td>
<td>85%</td>
<td>53%</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>Learning to operate will be easy for me</td>
<td>95%</td>
<td>90%</td>
<td>68%</td>
<td>0.01</td>
<td>87%</td>
<td>84%</td>
<td>95%</td>
<td>67%</td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Perceived Usefulness of Patient Portal Features by Age, Sex and Chronic Conditions

<table>
<thead>
<tr>
<th>Age (in years)</th>
<th>Sex</th>
<th>Schedule appointments</th>
<th>Request refills</th>
<th>View lab results</th>
<th>Review current meds</th>
<th>Manage medical issues</th>
<th>Email doctor re medical issues</th>
<th>Ask questions re medical issues</th>
<th>View screening tests</th>
<th>Get alerts/reminders</th>
<th>Share medical records with family</th>
<th>Share medical records with other doctors</th>
<th>View clinic notes</th>
<th>Communicate after hrs</th>
<th>Do office tasks online</th>
<th>Pre-appointment prep</th>
<th>Provide doctor w/ home bp or glucose reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 35 (n=43)</td>
<td></td>
<td>91%</td>
<td>95%</td>
<td>95%</td>
<td>95%</td>
<td>88%</td>
<td>91%</td>
<td>91%</td>
<td>95%</td>
<td>95%</td>
<td>47%</td>
<td>91%</td>
<td>91%</td>
<td>87%</td>
<td>77%</td>
<td>95%</td>
<td>83%</td>
</tr>
<tr>
<td>35-60 (n=43)</td>
<td></td>
<td>98%</td>
<td>93%</td>
<td>89%</td>
<td>91%</td>
<td>86%</td>
<td>91%</td>
<td>91%</td>
<td>95%</td>
<td>95%</td>
<td>51%</td>
<td>88%</td>
<td>83%</td>
<td>86%</td>
<td>81%</td>
<td>95%</td>
<td>86%</td>
</tr>
<tr>
<td>51+ (n=43)</td>
<td></td>
<td>84%</td>
<td>89%</td>
<td>91%</td>
<td>88%</td>
<td>86%</td>
<td>82%</td>
<td>79%</td>
<td>91%</td>
<td>96%</td>
<td>51%</td>
<td>86%</td>
<td>91%</td>
<td>90%</td>
<td>78%</td>
<td>95%</td>
<td>90%</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td>0.07</td>
<td>0.47</td>
<td>0.31</td>
<td>0.54</td>
<td>0.31</td>
<td>0.12</td>
<td>0.01</td>
<td>0.03</td>
<td>0.51</td>
<td>0.10</td>
<td>0.74</td>
<td>0.81</td>
<td>0.86</td>
<td>0.86</td>
<td>0.79</td>
<td>0.12</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td>96%</td>
<td>92%</td>
<td>92%</td>
<td>91%</td>
<td>92%</td>
<td>92%</td>
<td>91%</td>
<td>95%</td>
<td>95%</td>
<td>54%</td>
<td>88%</td>
<td>89%</td>
<td>88%</td>
<td>86%</td>
<td>87%</td>
<td>84%</td>
</tr>
<tr>
<td>Men (n=57)</td>
<td></td>
<td>92%</td>
<td>90%</td>
<td>92%</td>
<td>92%</td>
<td>92%</td>
<td>92%</td>
<td>91%</td>
<td>95%</td>
<td>95%</td>
<td>37%</td>
<td>91%</td>
<td>91%</td>
<td>91%</td>
<td>90%</td>
<td>93%</td>
<td>92%</td>
</tr>
<tr>
<td>Women (n=92)</td>
<td></td>
<td>80%</td>
<td>84%</td>
<td>87%</td>
<td>89%</td>
<td>90%</td>
<td>88%</td>
<td>84%</td>
<td>80%</td>
<td>87%</td>
<td>35%</td>
<td>90%</td>
<td>92%</td>
<td>91%</td>
<td>89%</td>
<td>90%</td>
<td>80%</td>
</tr>
</tbody>
</table>

### References
A Not So Lame Outlook for Injured Farm Workers: Return to Work Software Application Development

Bryan Weichelt, MS, MBA, PMP¹, Iris Anne Reyes, MPH¹, Andrea Mahnke, MS², William Ray, BS², Laurel Verhagen, BS², Shaun Halstead, BS², and Matthew Keifer, MD, MPH¹

¹National Farm Medicine Center, Marshfield Clinic Research Foundation, Marshfield, WI; ²Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI

Abstract

Serious, restrictive, non-fatal injuries are commonplace in large animal agriculture including in pork and dairy production. Primary care clinicians often have few resources to facilitate workers’ return to work. This project will develop a return to work software program to produce applicable light duty job assemblies (LDJA). One of the significant challenges is the integration of Physical and Occupational Therapists’ unstructured narrative data collection methods into structured data.

Introduction

Large animal farms are growing in size, increasing in work task specialization and hiring more immigrant labor. Workers not only face inherent risks in the agricultural workplace, but are also introduced to significant dangers in these operations. A range and variety of injuries in pork and dairy farms are common and are increasingly managed by primary care physicians. Yet, clinicians are often unfamiliar with the physical demands of farming and have little training and few resources to manage the return-to-work of injured workers.

The project utilizes examples of representative pork and dairy facilities to describe workflow, feasibility and impact of returning workers into the various LDJA positions. To create the LDJAs, the project is developing a computer application designed for clinicians, to guide early return-to-work planning for injured workers in the dairy and pork industries.

Methods

With assistance from the Marshfield Clinic Research Foundation’s Interactive Clinical Design Institute, we focused iterative design and prototype development on the needs of the users – clinicians, farm workers, and farm owners/managers. Focus groups of English and Spanish-speaking farm workers were conducted. One-on-one interviews with six different Marshfield Clinic clinicians led to significant design changes, shifting to a design that closely resembles the current Workers’ Compensation form used throughout the organization.

Physical and Occupational Therapists collected detailed functional job and task measurements on 32 implement dealers, dairies, and pork farms. The software application has several key components: 1) database of functional job profiles, 2) a clinician interface and input form, and 3) customized return-to-work output sheets for farm owners and their injured workers available in English or Spanish.

The functional job profile database consists of hazards and physical demands for common farm tasks collected by physical and occupational therapists. The clinician interface allows the provider to electronically enter restrictions. Algorithms automatically produce alternative job assemblies within the injured worker’s limitations. Lastly, simpler return-to-work output sheets are customized for owners and workers, replacing the traditional form.
Results

Physical and Occupational Therapist data collection is often based on a narrative and detailed approach. Limiting this input to structured elements allowed for algorithmic precision, but also hindered the collection and entry process, leading to four different input forms, including 24 versions of the field data collection form, which was developed in Microsoft Word.

Focus groups, interviews and usability tests with users were conducted throughout the development phase. Clinicians who were interviewed were aware of the usability issues of the complex return-to-work form currently used in practice but believed that the complexity was necessary. They expressed the desire to learn more about tasks on the farm through photos and videos, but felt that they had little time to do so.

Farm owners and workers preferred a simpler return-to-work form with lay person terms and would also like additional handouts regarding the injury or illness. All groups felt that the creation of a Spanish version of the form would be helpful for Spanish-speaking workers and their families.

Discussion

This project addresses an unmet need in agricultural safety and health – connecting the clinician to the farmer to reduce disability and sustain an adequate, safe workforce for the growing agriculture industry. It has also illuminated the significant challenge of addressing the unstructured data collection norms of Physical and Occupational Therapists in the field. The project identified those challenges, refined data collection and data entry tools and processes and developed new standards to accommodate.

Pruning descriptive value of agricultural tasks was an acceptable compromise to ensure a functional prototype was available for future testing in a clinical setting. To counter this, future research should include an evaluation of image files, .gif files, and video files as a replacement of descriptive narrative within the application.

This project closely links to several projects with a focus on injury and risk reduction in dairy production as part of the Upper Midwest Agricultural Safety and Health Center, whose overarching goal is to address health and safety issues faced by agriculture producers, workers, and their families in the Upper Midwest.
A Guideline for Assessing EHR Data Quality for Secondary Use
Nicole G. Weiskopf, PhD1, Chunhua Weng, PhD2
1Oregon Health & Science University, Portland, OR; 2Columbia University, New York, NY

Abstract
The reuse of electronic health record (EHR) data promises to increase the efficiency and generalizability of clinical research. EHR data, however, are known to be of poor and variable quality, which may decrease the validity of research findings. Researchers and informaticians require tools that allow them to select, perform, and report appropriate data quality assessments. We describe the formulation, development, and intended use of 3x3 DQA (Data Quality Assessment), a novel dynamic guideline for EHR data quality assessment.

Introduction
The secondary use of EHR data may circumvent some of the limitations of traditional prospective research, such as inefficiency and lack of generalizability.1-3 Unfortunately, it has been demonstrated that EHR data are generally of poor and highly variable quality.4-6 EHR data quality assessment must be transparent, task-dependent,7 systematic, evidence-based, and not reliant upon the availability of gold standard data. 3x3 DQA was designed to be responsive to the above principles. It is the first comprehensive guideline designed to mitigate and make transparent the effects of data quality on research validity through a systematic but task-dependent approach.

Methods
We conducted a series of studies to develop the guideline. Our approach combined qualitative studies of the attitudes and opinions of clinical researchers regarding their concerns about EHR data quality, systematic review of the informatics and data quality literature,8, 9 and quantitative analyses of EHR data quality.10 Through triangulation of the above, we derived a set of principles to guide contextual, task-dependent EHR data quality assessment.

Results
Version 1.0 of 3x3 DQA includes a user guide, questions to identify the data quality requirements of a study, a framework, and assessment and reporting recommendations. The guideline, which is intended to be used by informaticians and clinical researchers conducting clinical research with EHR-derived data, was developed for epidemiological study designs, though it may be extensible to other use cases.

The framework (Figure 1), which serves as the conceptual foundation of the guideline, features three constructs of data quality operationalized across three dimensions of the data. Because data quality is task-dependent, the guideline must be used for a specific dataset and associated research question(s). Once a study has been planned and a dataset has been extracted, a user steps through a series of questions to identify the relevant constructs, data dimensions, and assessment steps. This information then maps directly to the operationalizations (one per cell) in the framework. Each operationalization is in turn linked with a detailed methodological recommendation for assessment and reporting (Figure 2).

Discussion
Based on early expert feedback and user responses, 3x3 DQA has so far been well received by the EHR data quality community, though further work is warranted. A handful of projects using the guideline are underway. We are in the process of improving the clarity and usability of the guideline in response to user feedback. We intend to automate the guideline to reduce user burden and are planning a formal, scenario-based evaluation of the guideline.

Conclusion
3x3 DQA represents one of the first comprehensive methodological approaches to EHR data quality assessment. It is intended to enable researchers and informaticians engaged in the secondary use of EHR data to make determinations regarding the suitability of their data for their intended research and to conduct that research with greater confidence in the validity and trustworthiness of their results.
2. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93.
Making Hypertensive Medication Data Meaningful

Richard Williams, MA (Cantab)1,2, Benjamin Brown, MRCGP, MSc2, Niels Peek, MSc PhD1,2, Iain E. Buchan, MD, FACMI1,2

1 MRC Health eResearch Centre; 2 NIHR Greater Manchester Primary Care Patient Safety Translational Research Centre, University of Manchester, Manchester, UK.

Introduction

Primary care research databases derived from UK electronic health records (EHRs) contain coded prescription information. Much of this information is unstructured or unordered, and interpretation is required to assign clinical meaning to it. A common approach is to obtain the total mg of drug prescribed to a patient in a given time window and then calculate an average daily dose or exposure1-3. Another approach to inferring drug exposure is to count the number of prescriptions of the drug over the study period3-5, or in two adjacent time windows6. These approaches are valid if the dosage or presence of a drug is considered as a single covariate, however, to contextualize prescribing to the clinical decisions made on care pathways it is necessary to convert the EHR data into events, such as when therapy is commenced, terminated or changed. These events can then be used in a variety of analyses and tools such as: monitoring patient adherence to medication; care pathway analysis including process mining; next-generation phenotyping7; realistically-complex quality indicators; and advanced audit. Here we focus on medications prescribed for hypertension.

Method

We used an anonymized extract of primary care data from 53 general practices in Salford, UK (population 234k). The data contains Read codes (version 2) and EHR vendor-specific codes. All prescriptions of drugs recommended by the National Institute for Health and Care Excellence (NICE) for the treatment of hypertension8 in the UK were extracted. We created a mapping between each drug code, the active ingredient(s) and the tablet dose (mg). We found 653 Read codes and 199 vendor specific codes, covering 178 brand names and 70 generic names of antihypertensive medications. Text mining, using regular expressions, on 216,101 distinct textual patient instructions yielded the number of tablets taken per day. We then iteratively developed an algorithm to take the amount, frequency, duration and type of medication, together with the prescription date to extrapolate meaningful events as shown in Table 1 and Table 2. The iterations continued until the proportion of unclassified events achieved an acceptably low level. The algorithm was then validated by two authors (RW and BB, a clinician) who independently reviewed the records of a random sample of 100 patients to determine if the correct sequence of events had been extracted, and if not, recorded the discrepancy. Cohen’s κ showed fair inter-rater agreement (0.45). Disagreements were resolved through discussion.

Results

The algorithm was developed over 6 iterations involving the detailed examination of 179 patient records. A total of 10,311,973 prescriptions were extracted for 81,096 patients over the period 7 July 1977 to 12 December 2014 (54% female / median age 64, IQR [50, 77]). The breakdown for each family of drugs is shown in Table 3. The algorithm produced sequences with a combined total of 850,028 events (28% starts, 34% stops, 15% restarts, 16% increases, 8% decreases and 0.02% unclassified). During validation the algorithm achieved a PPV of 92% (95% CI 85%-96%). Of the 100 records reviewed only 8 had incorrect sequences and these were still partially correct. Four were missing a single stop event, two were false increases due to an erroneous prescription, one had an extra stop event, and one had a decrease for a switch from 2.5mg indapamide to 1.5mg modified release; these are actually clinically identical.

Discussion

We have developed a method for transforming unstructured prescription data into clinically meaningful events on a care pathway from EHR data. From these events it is easy to determine: when a patient is taking a medication; when there are adherence issues; when an intervention is or isn't made by a physician; when guidelines are being followed correctly; and if treatment is having the desired effect. We plan to incorporate these events into care pathways of chronic conditions in order to perform variance analysis to discover how variation from a pathway affects patient outcomes. The algorithm has the potential to be improved further by addressing the occasions where an incorrect sequence was produced. Strengths: The iterative nature of the design, high PPV, and clinically meaningful output. Limitations: We do not have information as to whether prescriptions were collected or taken. However, the act of prescribing is close to the clinician-centered decision context of care pathways.
Table 1. The extracted medication events and the reasoning for each one

<table>
<thead>
<tr>
<th>Medication event</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>Active ingredient first prescribed</td>
</tr>
<tr>
<td>Restart</td>
<td>Active ingredient previously prescribed but last event produced by the algorithm was a stop</td>
</tr>
<tr>
<td>Stop</td>
<td>Active ingredient previously prescribed but time has elapsed without a repeat prescription</td>
</tr>
<tr>
<td>Dose increase</td>
<td>Dose per day = (tablets per day) * (mg per tablet) increases</td>
</tr>
<tr>
<td>Dose decrease</td>
<td>Dose per day = (tablets per day) * (mg per tablet) decreases</td>
</tr>
</tbody>
</table>

Table 2. Example of a conversion from EHR to meaningful clinical events

<table>
<thead>
<tr>
<th>Date</th>
<th>EHR text</th>
<th>Instruction</th>
<th>Tabs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-02-15</td>
<td>Atenolol 25mg capsules</td>
<td>1 in the am</td>
<td>28</td>
</tr>
<tr>
<td>2014-03-16</td>
<td>Atenolol 25mg capsules</td>
<td>Take 2 daily</td>
<td>28</td>
</tr>
<tr>
<td>2014-04-01</td>
<td>Atenolol 50mg capsules</td>
<td>One each day</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 3. The number of antihypertensive drugs prescribed

<table>
<thead>
<tr>
<th>Drug family</th>
<th>Distinct drug types per family</th>
<th>Number of prescriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angiotensin-converting enzyme inhibitors</td>
<td>10</td>
<td>2,295,190</td>
</tr>
<tr>
<td>Angiotensin-II receptor blockers</td>
<td>7</td>
<td>841,217</td>
</tr>
<tr>
<td>Beta blockers</td>
<td>15</td>
<td>2,074,013</td>
</tr>
<tr>
<td>Calcium-channel blockers</td>
<td>11</td>
<td>2,021,822</td>
</tr>
<tr>
<td>Alpha blockers</td>
<td>8</td>
<td>543,226</td>
</tr>
<tr>
<td>Thiazide-like diuretic</td>
<td>11</td>
<td>1,278,005</td>
</tr>
<tr>
<td>Other diuretics</td>
<td>8</td>
<td>1,258,500</td>
</tr>
</tbody>
</table>

Acknowledgements: Funded by the National Institute for Health Research Greater Manchester Primary Care Patient Safety Translational Research Centre (NIHR GM PSTRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Transparency: The algorithm and scripts used here will be made publically available on https://github.com/ and the code lists uploaded to http://clinicalcodes.org.

References

Improving Weight in Patients with Serious Mental Illness: A Randomized Controlled Trial of Computerized Weight Services with Peer Coaches

Alexander S. Young, MD, MSHS1,2, Amy N. Cohen, PhD1,2, Richard Goldberg, PhD3
Julie Kreyenbuhl, PharmD3, Fiona Whelan, MA2
1Department of Veterans Affairs, Los Angeles, CA, USA; 2UCLA, Los Angeles, CA, USA; 3Department of Veterans Affairs, Baltimore, MA, USA

Abstract

In-person weight management interventions improve outcomes of people with serious mental illness. However, these require extensive time from patients and clinicians, and are rarely provided in usual care. We studied whether barriers could be addressed using computerized provision of diet and exercise education and decision support, combined with motivation from peer coaches.

Introduction

People with serious mental illness are at high risk for obesity and related medical problems, and die 10 to 20 years prematurely, most commonly from cardiovascular disease. Guidelines recommend specialized, in-person weight management interventions. These result in weight loss when delivered in efficacy trials done with motivated patients who are often paid to participate. In usual practice, these interventions are rarely provided, and, when provided, patient enrollment and retention are low. Interventions require substantial clinician time and frequent clinic visits for patients. We studied whether these barriers could be addressed using computerized provision of diet and exercise education and decision support, combined with motivation and support from peer coaches.

Methods

Randomized, controlled, comparative effectiveness study, conducted with overweight patients with serious mental illness at mental health clinics. 276 patients were recruited from a Veterans Affairs medical center, and 54 patients recruited from a county mental health clinic. Patients were randomized to 1) online weight management with peer coaching, 2) in-person clinician-led weight services, or 3) to continue with treatment as usual. Online weight management included 30 modules plus weekly telephonic peer coaching. The online system could be accessed from clinic kiosks, or anywhere there is internet access. It provided simultaneous audio and text-based education, video, pedometer tracking, goal setting, homework, diet plans, and quizzes. Coaching was delivered by individuals with lived experience with mental illness, was phone-based, and utilized motivational interviewing principles. In-person weight management had the same curriculum as the online program. At 6 months, patient outcomes were assessed.

Results

A mixed measures repeated model predicted Body Mass Index (BMI) at 6 months while controlling for site and BMI 6 months prior to entering the study. 271 individuals were obese at baseline (BMI > 30). In obese patients, there was a significant group by time interaction (F=3.4, p=.03). The online and peer coaching group had weight loss averaging 0.8 ± 0.2 BMI points (5 pounds; p<.01). No change was seen in treatment as usual (p=.42) or in-person services (p=.57). No effect was seen in patients with BMI < 30. 21% of patients completed the on-line program compared to 0% completing all in-person groups (χ²=19.7; p<.0001). The on-line intervention was well received.

Discussion

On-line weight management with peer supports can provide educational content and decision support that is tailored to individual patients, and patient-centered. It is feasible in usual care settings, has little burden on clinicians, and marginal costs are low. It produces weight loss, and may have greater effectiveness than clinician-led services. Patient motivation remains a critical challenge to spreading weight interventions beyond efficacy trials in highly select samples. Informatics technology and peer supports show the potential to enhance motivation, though continued innovation can strengthen mean outcomes in usual care populations.

Reference

Abstract Title: Iterative design and evaluation methodology for clinical decision support systems

Authors: Fei Yu, PhD, Vincent Carrasco, MD, Ketan K. Mane, PhD, Javed Mostafa, PhD

Introduction: Wide adoption of CDS systems among physicians remains an elusive goal. The MindsEye project was initiated as a collaboration among UNC informatics researchers and clinicians to respond to this need. The goal of this project was to provide an interactive user interface (UI) that: organizes and presents data by visualization; allows sorting, sifting, and re-organization of data on demand and seamlessly integrates with healthcare practice workflow, as well as evaluating a new UI for the treatment of outpatient psychiatric patients (Figure 1).

Methods: An iterative development and evaluation methodology was adopted and implemented as a two-phase strategy. Phase I concentrated on design, evaluation and refinement. Pilot users (n=6) provided feedback on a range of tasks. This data was carefully aggregated, analyzed and converted to UI refinements. Subsequently, a more comprehensive study was conducted, with the refined interface, as Phase II (n=19). Both phases consisted of a 4-step usability process: pre-test questionnaire, UI orientation session, monitored task performance session, and post-test questionnaire (Table 1). The monitored task performance stage used a think-aloud protocol during the task sessions. The subjects’ opinions, comments, and feelings were recorded. At the end of the post-test survey, a discussion session was held with each subject to clarify the collected comments and feedback. A video-based screen capturing software, called Camtasia, was used to record all task sessions for analysis and evaluation.

Results: Data from Phase I and Phase II were compared (Tables 2, 3, and 4). Comparatively, Phase II subjects demonstrated improved performance and satisfaction. On critical features, the results were consistent across both phases. For example, Timeline view was consistently rated highest on like-ability whereas the treatment guideline feature was poorly rated in both phases. The forecasting section yielded inconsistent results between two phases. It was generally observed that the like-ability factor and the subject task performance did not correlate in either phase.

Discussion: Overall, this iterative UI design and evaluation methodology proved useful and produced superior user performance and satisfaction on critical features of the refined version of the UI. Among all the UI features, physicians consistently rated Timeline view the best. However, they requested further refinements. Another important lesson we learned was that design and evaluation techniques require flexibility to accommodate the dynamic clinical ecosystem.

References

Figures
Figure 1. Screen shots of the MindsEye UI: Phase I vs. II
Table 1. The 4-step usability testing process

<table>
<thead>
<tr>
<th>#</th>
<th>Step</th>
<th>Data Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pre-test questionnaire</td>
<td>• Demographic info (e.g., age, gender)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Primary area of clinical practice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Professional experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• EHR use experience and frequency</td>
</tr>
<tr>
<td>2</td>
<td>UI orientation (5-minute overview)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Task performance</td>
<td>Each physician subject was given a set of tasks to complete (Table 2). All task sessions were recorded.</td>
</tr>
<tr>
<td>4</td>
<td>Post-test questionnaire</td>
<td>Answers to 16 user survey questions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Ease of learning &amp; helpfulness of UI features</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• User perception on UI design (color scheme, texts display, recovery from mistakes, etc.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• User satisfaction with UI</td>
</tr>
</tbody>
</table>

Table 2. Usability task comparison: Phase I vs. II

<table>
<thead>
<tr>
<th>TASK #</th>
<th>Phase I</th>
<th>Data view needed to complete task</th>
<th>Phase II</th>
<th>Data view needed to complete task</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Timeline</td>
<td></td>
<td>Patient Demographics</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Patient Rx Profile</td>
<td></td>
<td>Comorbidity</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Rx Recommendation</td>
<td></td>
<td>Patient Rx Profile</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Treatment Guideline</td>
<td></td>
<td>Timeline</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Treatment Guideline</td>
<td></td>
<td>Treatment Guideline</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Filtering</td>
<td></td>
<td>Rx Recommendation &amp; Predictive</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Predictive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Task performance comparison: Phase I vs. II

<table>
<thead>
<tr>
<th>TASK #</th>
<th>Phase I</th>
<th>Data view needed to complete task</th>
<th>Out of 6 subjects, how many got it correct (percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Timeline</td>
<td>1 (17%)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Patient Rx profile</td>
<td>4 (67%)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Rx recommendation</td>
<td>6 (100%)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Treatment guideline</td>
<td>4 (67%)</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Treatment guideline</td>
<td>5 (83%)</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Filtering</td>
<td>5 (83%)</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Predictive</td>
<td>5 (83%)</td>
</tr>
</tbody>
</table>

Table 4. Efficiency and like-ability factor comparison: Phase I vs. II

<table>
<thead>
<tr>
<th>Data View</th>
<th>Phase I</th>
<th>Efficiency in correctly communicating the info</th>
<th>Like-ability Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timeline</td>
<td></td>
<td>17% (1/6)</td>
<td>★★★★</td>
</tr>
<tr>
<td>Patient Rx Profile</td>
<td></td>
<td>68%</td>
<td>★★★★</td>
</tr>
<tr>
<td>Rx Recommendation</td>
<td></td>
<td>100%</td>
<td>★★★★</td>
</tr>
<tr>
<td>Treatment Guideline</td>
<td></td>
<td>68% - 85%</td>
<td>★★★★</td>
</tr>
<tr>
<td>Filtering</td>
<td></td>
<td>85%</td>
<td>★★★★</td>
</tr>
<tr>
<td>Predictive Insight</td>
<td></td>
<td>85%</td>
<td>★★★★</td>
</tr>
<tr>
<td>Data View</td>
<td>Phase II</td>
<td>Efficiency in correctly communicating the info</td>
<td>Like-ability Factor</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------</td>
<td>------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Patient Demographics</td>
<td></td>
<td>100%</td>
<td>N/A</td>
</tr>
<tr>
<td>Comorbidity</td>
<td></td>
<td>95%</td>
<td>N/A</td>
</tr>
<tr>
<td>Patient Rx Profile</td>
<td></td>
<td>95%</td>
<td>★★★★</td>
</tr>
<tr>
<td>Timeline</td>
<td></td>
<td>95%</td>
<td>★★★★</td>
</tr>
<tr>
<td>Treatment Guideline</td>
<td></td>
<td>68% - 89%</td>
<td>★★★★</td>
</tr>
<tr>
<td>Rx Recommendation &amp; Predictive View</td>
<td></td>
<td>95%</td>
<td>★★★★</td>
</tr>
</tbody>
</table>
Improving Retrieval of PubMed Articles Using the TopicalMeSH Representation

Zhiguo Yu, MS¹, Elmer Bernstam, MD, MSE¹, Trevor Cohen, MBChB, PhD¹, Byron C. Wallace, PhD², Todd R. Johnson, PhD¹
¹The University of Texas School of Biomedical Informatics at Houston, Houston, TX; ²School of Information, University of Texas at Austin, Austin, TX

Introduction
Medical Subject Headings (MeSH) was developed by the US National Library of Medicine (NLM) to better manage and search large volumes of articles in MEDLINE. But MeSH has several limitations, including: the lack of concept coverage for newly developing areas; human inconsistency in assigning codes; and the time required to manually index an exponentially growing corpus. In contrast, probabilistic topic modeling approaches can automatically index and summarize corpora. An important practical question is whether automatically generated topics have any advantage over MeSH indexing alone. To address this question, we introduced the TopicalMeSH representation for biomedical literature by leveraging the correspondence between latent topics (uncovered via topic modeling) and MeSH. We evaluated TopicalMeSH as a representation for document retrieval and classification on a corpus comprising 15 drug reviews. TopicalMeSH performed better than MeSH alone in both of these tasks.

Method
Chuang et al. developed a topic model evaluation framework by computing a correspondence matrix of reference topics assigned by domain experts and latent topics recovered by Latent Dirichlet Allocation (LDA). LDA is a generative model that considers each document to be a mixture of latent topics, and defines these topics as distributions over words. We propose a variation of this method to compute the correspondence between MeSH terms and topics generated by LDA. To compute the similarity score between inferred topics and MeSH, we represent each MeSH term as a distribution of the words contained in documents that were tagged with that MeSH term. Figure 1 depicts our approach schematically. Compute Correspondence Matrix uses the rescaled dot product defined below to calculate the similarity score of each MeSH term-topic pair. $M_5$ is the Topic-to-MeSH correspondence matrix. $M_6$ is our document-to-TopicalMeSH matrix, which is the matrix product of $M_5 \times M_4$. This matrix relates MeSH terms to documents in which corresponding topics feature prominently. We used two tasks to evaluate the utility of the TopicalMeSH feature matrix $M_6$ and the original MeSH features in Matrix $M_1$.

\[
\text{Re} \text{scaled Dot Product} = \frac{\hat{p}_Q - d_{\text{ms}}}{d_{\text{ms}} - d_{\text{ms}}}; \quad d_{\text{ms}} = \frac{\hat{p}Q}{\hat{q}}; \quad d_{\text{ms}} = \frac{\hat{p}Q}{\hat{q}}; \quad \hat{p} \text{ and } \hat{q} \text{ are vectors consisting of all unique words in descending order}; \quad \hat{q} \text{ is in ascending order.}
\]

The first task was a document retrieval task. We mapped the user's information needs (for a systematic review) to candidate relevant MeSH terms. We used these MeSH terms, both directly and via TopicalMeSH, to rank documents and then calculated the Mean Average Precision (MAP) achieved using these representations. The second evaluation was a classification task, in which we applied a supervised machine learning method to the corpus to learn to classify documents with respect to their relevance. We used Support Vector Machines (SVMs) to compare the performance ($F_1 =$ harmonic mean of recall and precision) of the following five different representations: MeSH, TopicalMeSH, Words, MeSH+Words, TopicalMeSH+Words. For both tasks, we used 15 systematic drug review corpora reported by Cohen et al.

Results
Table 1 shows the results of the document retrieval task. TopicalMeSH achieved a higher MAP score in 11 of 15 corpora. The average MAP improvement from MeSH to TopicalMeSH was about 5%. For the classification task, Table 2 shows the results for SVMs that performed best. Again, TopicalMeSH achieved a higher $F_1$ in 14 of 15 corpora compared to MeSH. TopicalMeSH+Words had a higher $F_1$ than MeSH+Words in 11 of 15 corpora (Table 2).

Conclusion
The TopicalMeSH representation we propose, which combines latent topics and corresponding MeSH terms, improves performance on information retrieval and document classification tasks involving PubMed abstracts.
References


Figure 1. A schematic of our approach to inducing TopicalMeSH representations of articles

Table 1. Mean Average Precision results of the document retrieval task.

<table>
<thead>
<tr>
<th>Drug Review Name</th>
<th>MeSH</th>
<th>TopicalMeSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>7.51%</td>
<td>10.95%</td>
</tr>
<tr>
<td>ADHD</td>
<td>47.77%</td>
<td>27.4%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>40.64%</td>
<td>31.7%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>33.56%</td>
<td>43.35%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>16.71%</td>
<td>18.26%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>25.64%</td>
<td>31.28%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>29.89%</td>
<td>37.74%</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>23.06%</td>
<td>41.54%</td>
</tr>
<tr>
<td>Opioids</td>
<td>2.86%</td>
<td>2.72%</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>27.07%</td>
<td>43.83%</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>20.25%</td>
<td>20.4%</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>3.9%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Statins</td>
<td>5.57%</td>
<td>11.46%</td>
</tr>
<tr>
<td>Triptans</td>
<td>41.07%</td>
<td>50.75%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>19.66%</td>
<td>44.93%</td>
</tr>
</tbody>
</table>

EPC = Evidence-Based Practice Center; ACE = angiotensin-converting enzyme; ADHD = attention deficit/hyperactivity disorder; NSAIDs = nonsteroidal anti-inflammatory drug.

Table 2. F1 of Support Vector Machines in the classification task.

<table>
<thead>
<tr>
<th>Drug Review Name</th>
<th>MeSH</th>
<th>TopicalMeSH</th>
<th>Words</th>
<th>MeSH+Words</th>
<th>TopicalMeSH+Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>7.51%</td>
<td>10.95%</td>
<td>20%</td>
<td>17.7%</td>
<td>25.7%</td>
</tr>
<tr>
<td>ADHD</td>
<td>47.77%</td>
<td>27.4%</td>
<td>30.5%</td>
<td>53.7%</td>
<td>49.6%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>40.64%</td>
<td>31.7%</td>
<td>27.1%</td>
<td>53.1%</td>
<td>35.5%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>33.56%</td>
<td>43.35%</td>
<td>53.7%</td>
<td>44.5%</td>
<td>56.2%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>16.71%</td>
<td>18.26%</td>
<td>32.2%</td>
<td>29.6%</td>
<td>36.7%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>25.64%</td>
<td>31.28%</td>
<td>43.8%</td>
<td>36.5%</td>
<td>50.5%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>29.89%</td>
<td>37.74%</td>
<td>22.8%</td>
<td>29%</td>
<td>36.6%</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>23.06%</td>
<td>41.54%</td>
<td>46.3%</td>
<td>36.3%</td>
<td>50.9%</td>
</tr>
<tr>
<td>Opioids</td>
<td>2.86%</td>
<td>2.72%</td>
<td>4%</td>
<td>10.7%</td>
<td>4%</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>27.07%</td>
<td>43.83%</td>
<td>29.2%</td>
<td>45.3%</td>
<td>48.6%</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>20.25%</td>
<td>20.4%</td>
<td>33.6%</td>
<td>26.4%</td>
<td>37.5%</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>3.9%</td>
<td>3.4%</td>
<td>26.3%</td>
<td>0</td>
<td>29.8%</td>
</tr>
<tr>
<td>Statins</td>
<td>5.57%</td>
<td>11.46%</td>
<td>18.2%</td>
<td>12.4%</td>
<td>19%</td>
</tr>
<tr>
<td>Triptans</td>
<td>41.07%</td>
<td>50.75%</td>
<td>63.3%</td>
<td>59.6%</td>
<td>64.9%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>19.66%</td>
<td>44.93%</td>
<td>12%</td>
<td>52.57%</td>
<td>33.63%</td>
</tr>
</tbody>
</table>
Towards Computational Drug Repositioning: A Comparative Study of Single- and Multi-task Learning

Ping Zhang, PhD\(^1\), Zhaonan Sun, PhD\(^1\), Fei Wang, PhD\(^2\), Jianying Hu, PhD\(^1\)
\(^1\)IBM T. J. Watson Research Center, Yorktown Height, NY; \(^2\)University of Connecticut, Storrs, CT

Introduction

In response to the high cost and high risk associated with traditional de novo drug discovery, investigation of potential additional uses for existing drugs, also known as drug repositioning, has attracted increasing attention from both the pharmaceutical industry and the research community. We hypothesize that diseases are related to each other, and learning multiple related tasks simultaneously could improve the prediction performance of drug repositioning. In this study, we propose a multi-task learning (MTL) framework to generate drug repositioning hypotheses and compare it with single-task learning (STL) strategy.

Methods

We collected 1255 drugs from DrugBank. Each drug was represented by an 881-dimensional binary profile whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. The known uses of these drugs were collected from MEDI database \(^1\). Indications in MEDI were coded as ICD9 codes. We grouped ICD9 codes based on their first 3 digits to avoid trivial predictions (i.e., repurpose a drug from a disease to very similar diseases). Also, we excluded non-disease conditions and rare diseases, and obtained 100 ICD9 groups as diseases in our drug repositioning experiment. Between our 1255 drugs and 100 diseases, there are 3430 distinct drug-disease interactions in the dataset.

We modeled the drug repositioning task as a binary classification problem. We constructed a classifier for predicting whether a given drug could treat a particular disease or not, and repeated this process for all 100 diseases. In the scenario of STL, such 100 tasks are solved independently, ignoring the task relatedness. However, the tasks of predicting disease indications of drugs are related (as a disease might associate with other diseases in real life). In MTL, these related tasks are learnt simultaneously by extracting and utilizing appropriate shared information across tasks. Figure 1 illustrates the difference between STL and MTL.

In our study, we used logistic regression as the base classifier to compare STL and MTL. We extended a multi-task relationship learning \(^2\) model to deal with binary targets and applied it to drug repositioning problem. We also investigated whether imposing a prior to target relatedness could improve the performance of MTL predictions. In our problem, the prior knowledge comes from the hierarchical structure of ICD9 codes (i.e., disease areas and grouping information of the targets). In total, we considered three drug repositioning methods in the experiment: (1) STL; (2) MTL without prior; (3) MTL with prior. We used a 10-fold cross validation scheme to evaluate the performance of all methods.

Results

Figure 2 shows the ROC curves of the three methods in the experiment. The figure shows that MTL without prior and MTL with prior obtained AUC scores of 0.8431 and 0.8592 respectively, which are significantly higher than that of STL (0.7930). In our drug repositioning experiment, learning multiple related tasks simultaneously effectively increases the training sample (i.e., known drug-disease interactions) size for each task, and improves the prediction performance. Figure 2 also shows that incorporating a prior from ICD9 hierarchical structure could improve the performance of MTL predictions.

As an important by-product, MTLs also learn target relatedness (i.e., disease associations) during the predictions. Figure 3 is a heatmap of learned target relatedness from MTL with prior. In the heatmap, a high value between two tasks indicates the two diseases are highly correlated during the predictions. For example, ICD9 group 401 (Hypertension) is found to be highly correlated with ICD9 group 299 (Autism) by MTL with prior, which is in agreement with a recent clinical trial \(^3\). The learned target relatedness from MTL could provide additional insights for downstream investigations including clinical trials.
Discussion

To our knowledge, the MTL framework proposed in the study is the first feature-vector based multi-task learning method for drug repositioning hypothesis generation. Our preliminary results demonstrate that learning multiple related tasks simultaneously is effective in achieving improved performance. Furthermore, learned target relatedness from MTL could help to reveal MoA of drug repositioning hypotheses. In future work, we intend to integrate more features of drugs (e.g., targets, side effects) to the predictions, and to incorporate and compare more prior information (e.g., disease co-occurrence in real-world data) on initializing task relatedness.

\[
y \in \mathbb{R}^N, \quad X \in \mathbb{R}^{N \times D}, \quad w \in \mathbb{R}^D, \quad Y \in \mathbb{R}^{N \times S}, \quad X \in \mathbb{R}^{N \times D}, \quad W \in \mathbb{R}^{D \times S}
\]

(a) Single-task learning

(b) Multi-task learning

**Figure 1.** Illustration of (a) single-task learning (STL) and (b) multi-task learning (MTL). In STL, each task is considered to be independent and learnt independently. In MTL, multiple tasks are learnt simultaneously, by utilizing task relatedness.

**Figure 2.** The ROC comparison of three methods on drug repositioning experiment.

**Figure 3.** Heatmap of learned disease relatedness from MTL with prior model.

References


Identification of Venous Thromboembolism from Electronic Medical Records with Information Extraction

Shuai Zheng, PhD¹, Raymund Dantes MD, MPH¹, James J. Lu, PhD¹, Sheri Chernetsky Tejedor, MD¹, Michele Beckman, MPH², Asha Krishnaswamy, MSc², Lisa C. Richardson, MD², Fusheng Wang, PhD¹

¹ Emory University, Atlanta, GA
² Centers for Disease Control and Prevention, Atlanta, GA

Introduction:

Venous thromboembolism (VTE), including deep vein thrombosis (DVT) and pulmonary embolism (PE), is associated with significant morbidity and mortality, resulting in approximately 100,000 deaths per year in the U.S [1]. VTE can be diagnosed by several radiologic studies, including lower or upper extremity ultrasonography and computerized tomography (CT) of the chest. Federally mandated reporting of VTE defined by the Agency for Healthcare Research and Quality Patient Safety Indicator 12 (AHRQ PSI-12) [2] is based on administrative and billing data, whose accuracy for detecting VTE has yet to be demonstrated. While automated information extraction (IE) from electronic medical records has high potential for improving VTE identification, existing tools either require large, manually annotated datasets for training, or considerable linguistic expertise for constructing domain specific rules. We present IDEAL-X [3], a novel open source IE system, and evaluate its accuracy for identifying VTE diagnosis directly from radiology reports in electronic medical records.

Methods:

Full text of radiology reports and clinical data were extracted from the electronic medical records (Cerner Corp, Kansas City, MO) of 13,248 patients admitted to Emory University Orthopedic and Spine Hospital from 2009-2014. Patient encounters were defined as a hospital admission where both surgery (of the spine, hip, or knee) and a radiology diagnostic study for VTE were performed. A physician manually reviewed each radiology report for diagnosis of a DVT or PE. The IDEAL-X system analyzed the same radiology report under two separate modes: i) controlled vocabulary mode, where the user specifies upfront terminology and contextual information (such as relevant and irrelevant report sections) to be extracted, and ii) online machine learning mode, where all terminology and contextual information is learned incrementally. The system processes a report and proposes VTE diagnostic values for a user to review. The user’s feedback (accepting or correcting system proposed values) provides linguistic features of the value extracted, which automatically update and improve the system’s extraction model for the next iteration. Performance was analyzed for total radiology reports, and patient encounters (multiple reports per encounter possible).

Results:

Among 2083 radiology reports in the testing dataset, IDEAL-X in controlled vocabulary mode correctly identified 176/181 VTE events, achieving a sensitivity of 97.2% (95% Confidence Interval [CI] 93.7-99.1%) and specificity of 99.3% (95% CI 98.9-99.7%) when compared to manual review (Table 1). This performance was superior to online machine learning mode, which achieved an overall sensitivity of 92% (95% CI 88.3%-96.1) and 99% specificity (95% CI 98.5%-99.4%), and required approximately 50% of reports to be processed before achieving >95% sensitivity and specificity (Figure 1). Among 422 surgical encounters with diagnostic radiographic studies for VTE, IDEAL-X in controlled vocabulary mode correctly identified 41/42 VTE events, achieving a sensitivity of 97.6% (95% CI 97.4-97.8%) and specificity of 99.8% (95% CI 98.7-100.0%) (Table 2). The performance surpasses that of AHRQ-PSI 12 [2], which has sensitivity of 92.9% (95% CI 80.5-98.4%) and specificity of 92.9% (95% CI 89.8-95.3%), though only the difference in specificity was statistically significant (p<0.01).

Discussion:

IDEAL-X is able to correctly identify VTE from the free text of radiology reports with very high sensitivity and specificity, surpassing the performance of identification based on AHRQ PSI-12. Clinical quality metrics sourced from clinical records may have increased validity compared to those from administrative data sources. Customized controlled vocabulary simplifies the deployment process, and is better suited for instances with low positive incidences, including VTE. Online machine learning enables a user to provide real-time feedback to continually improve the accuracy of extraction of subsequent reports. IDEAL-X’s workflow requires no linguistic expertise from the user, and can be easily adapted to different clinical applications to improve detection and surveillance of medical conditions. A comparative study of IDEAL-X against other natural language processing systems for DVT detection is beyond the space available here, and will be reported in the future.
Table 1. Performance of IDEAL-X in Controlled Vocabulary Mode, Analyzing Total Radiology Reports

<table>
<thead>
<tr>
<th>Event</th>
<th>Radiology Report Types</th>
<th>Total Reports</th>
<th>Positive Reports By Manual Review</th>
<th>Positive Reports By IDEAL-X</th>
<th>Measure</th>
<th>IDEAL-X Performance (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DVT</td>
<td>Ultrasonography of Upper or Lower Extremity</td>
<td>1153</td>
<td>112</td>
<td>109</td>
<td>Sensitivity</td>
<td>97.3% (92.4-99.4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Specificity</td>
<td>99.4% (98.7-99.8%)</td>
</tr>
<tr>
<td>PE</td>
<td>CT and MRI of Chest</td>
<td>930</td>
<td>69</td>
<td>67</td>
<td>Sensitivity</td>
<td>97.1% (89.9-99.6%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Specificity</td>
<td>99.3% (98.5-99.7%)</td>
</tr>
<tr>
<td>Either DVT or PE</td>
<td>All four types above</td>
<td>2083</td>
<td>181</td>
<td>176</td>
<td>Sensitivity</td>
<td>97.2% (93.7-99.1%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Specificity</td>
<td>99.3% (98.9-99.7%)</td>
</tr>
</tbody>
</table>

Table 2. Performance of IDEAL-X System in Controlled Vocabulary Mode, compared to Agency for Healthcare Research and Quality Patient Safety Indicator 12, Analyzed by Patient Surgical Encounter

<table>
<thead>
<tr>
<th>Event</th>
<th>Total Patients</th>
<th>Events by Manual Review</th>
<th>Events by IDEAL-X</th>
<th>Events by AHRQ-PSI 12</th>
<th>Measure</th>
<th>IDEAL-X (95% CI)</th>
<th>AHRQ-PSI 12 (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DVT</td>
<td>422</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>Sensitivity</td>
<td>94.1% (71.2-99.0%)</td>
<td>76.5% (50.1-93.0%)</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Specificity</td>
<td>100.0% (99.1-100.0%)</td>
<td>96.1% (93.7-97.7%)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>PE</td>
<td>422</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>Sensitivity</td>
<td>100.0% (86.2-100.0%)</td>
<td>100.0% (86.16-100.0%)</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Specificity</td>
<td>99.8% (98.6-100.0%)</td>
<td>95.7% (93.9-97.5%)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Either DVT or PE</td>
<td>422</td>
<td>42</td>
<td>41</td>
<td>39</td>
<td>Sensitivity</td>
<td>97.6% (87.4-99.6%)</td>
<td>92.9% (80.5-98.4%)</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Specificity</td>
<td>99.8% (98.7-100.0%)</td>
<td>92.9% (89.8-95.3%)</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Figure 1. Rates of Sensitivity and Specificity Improvement over Processed Records

References
Rising Drug Allergy Alert Overrides in a Computerized Provider Order Entry System: a Decade of Experience

Li Zhou MD, PhD1,2, Maxim Topaz PhD, RN1,2, Diane L Seger RPh1,3, Sarah P Slight PhD, PGDip1,4, Foster Goss DO4, Kenneth Lai MA3, Paige G Wickner MD, MPH1, Kimberly Blumenthal MD1,6, Neil Dhapeshwarkar BA1, Frank Chang MSE3, David W Bates MD1,2

1Brigham and Women's Hospital, Boston, MA, USA; 2Harvard Medical School, Boston, MA, USA; 3Partners HealthCare System, Boston, MA; 4School of Medicines, Pharmacy and Health, Durham University, Durham, UK; 5University of Colorado, Aurora, CO, USA; 6Department of Medicine, Massachusetts General Hospital, Boston MA, USA

Introduction

The electronic health record (EHR) with computerized provider order entry (CPOE) can provide the clinician with real-time guidance and support at the point of prescribing1. These systems have been shown to improve the safety and quality of patient care by, for example, reducing the likelihood of medication errors and consequent patient harm2. One of the basic drug interaction alert types is drug allergy checking. This provides the clinicians with an alert or a reminder if the patient has a documented allergy to the prescribed drug in their EHR. However, clinicians are often exposed to a high number of drug allergy interaction (DAI) alerts, which can result in them experiencing alert fatigue; Consequently, clinicians can ignore or override both clinically important and unimportant alerts3,4. Ignoring alerts can potentially lead to patient harm and other unintended consequences, thus many efforts are underway to improve the accuracy of the alerts and reduce alert fatigue5,6. Unfortunately, little is known about the acceptance rates and other aspects of DAI alerts presented to providers.

In this study, we aimed to: 1) evaluate providers’ DAI alert overrides over the last ten years (2004-2013); 2) examine the different types of reactions overridden and whether these related to definite, probable or possible drug-allergy matches; and 3) examine the reasons why providers chose to override these drug-allergy interaction alerts.

Methods

In this observational, cross sectional study, we assessed ten years of DAI alert records from two large academic medical centers (Brigham and Women's Hospital and Massachusetts General Hospital). Both hospitals are part of Partners Healthcare, an integrated healthcare system in the Boston area. Data were extracted from the Partners Enterprise-wide Allergy Repository (PEAR), a longitudinal allergy database shared within the Partners provider/hospital network7. We identified all DAI alerts (n=928,962) from the PEAR database between Jan 2004 and Dec 2013. We assessed the most commonly triggered DAI alerts with frequencies higher than 0.1% of total alerts, resulting in 611,192 records (65.8%) included in this analysis.

In PEAR, medications are stored and encoded using a combination of proprietary (First Databank, Inc.™) and local (Partners Master Drug Dictionary) terminologies. For patients admitted to the healthcare system, providers enter the patient’s allergy information in the EHR. The system then checks every prescribed medication against possible allergies in PEAR. Drug-allergy interaction alerts are triggered when a prescribed medication matches the stored allergy as either: Definite match (e.g., levofloxacin -> levofloxacin): where there is an exact match between the allergen and prescribed medication (or main medication ingredient); Probable match (e.g., ciprofloxacin -- > levofloxacin): where the prescribed medication matches the allergen group of the documented allergen; or Possible match (e.g., penicillin (penicillins) -- > cephalexin (cephalosporins): where the cross-sensitivity group of the patient’s allergen matches the cross-sensitivity group of the medication ingredient. Each alert warning lists one or more reactions that the patient is at risk of being exposed to (e.g., anaphylaxis). We reviewed the different types of reactions as well as the current literature8,9 and classified them into two main types: (1) Potentially immune mediated (e.g., rash or anaphylaxis) or non-immune mediated (e.g., nausea or vomiting), and (2) Potentially life threatening (e.g., anaphylaxis or bronchospasm) or non-life threatening (e.g., gastrointestinal upset or nausea).
**Results**

Overall, we found an increase in the rate of DAI alert overrides, from 83.3% in 2004 to 87.6% in 2013 (p<.001). The most common DAI alerts were triggered by allergies to narcotics (48%) and other analgesics (6%), antibiotics (10%) and statins (2%). Only slightly more than one-third (35.1%) of drugs overridden had reactions that were potentially immune mediated. Providers were least likely to override alerts for definite allergen and prescribed medication matches that resulted in an immune mediated (72.8%) and life threatening reactions (74.1%) as compared to possible (cross-sensitivity) or probable (allergen group) matches. Finally, more than half of the overrides reasons pointed to irrelevant alerts (i.e. Patient has tolerated the medication before, 50.9%) and providers were significantly more likely to override repeated alerts (89.7%) rather than first time alerts (77.4%, p < .001).

**Limitations**

Our data were confined to two large academic hospitals in Boston, which limits the generalizability of our findings. In addition, although based on the literature and expert panel discussions, our classification of potentially life threatening and potentially immune mediated reaction is an estimation based on the captured patient reactions rather than a conclusive statement based on patient lab/pathology or allergy tests.

**Conclusion**

We examined one of the largest reported allergy datasets for trends in the most commonly triggered drug allergy interaction alerts in the inpatient setting. The most common DAI alerts were triggered by allergies to narcotics and other analgesics, several antibiotic groups, and statins. Only slightly more than one-third of the reactions were potentially immune mediated. Over the past decade, we identified an alarming trend of increasing rate of DAI alert overrides, culminating in 87.6% in 2013. The DAI alert override rates were similarly high for both potentially immune mediated or life threatening reactions. Overall, definite allergy and prescribed medication matches were overridden less frequently compared to probable or possible matches. The most overridden medication groups were the narcotic analgesics and statins. Finally, more than half of the overrides reasons pointed to irrelevant alerts and providers were significantly more likely to override repeated alerts rather than first time alerts. Our findings underline the urgent need for more efforts focused on providing more accurate and relevant drug allergy interaction alerts for providers to decrease alert overrides and alert fatigue. Finally, further investigation is needed into providers’ reasons for overriding DAI alerts.

**References**

8. Burks AB, W; Holgate, ST; Lemanske, RF; O’Hehir, RE; Adkinson, NF; Bochner, BS Middleton's Allergy: Principles and Practice. 8 ed: Elsevier Health Services; 2013.
The sociotechnical perspective in biomedical informatics: what do we understand and measure?

Panelists
Jos Aarts, PhD, FACMI1, Joan Ash, PhD, FACMI2, Andre Kushniruk, PhD, FACMI3, Dean F. Sittig, PhD, FACMI4

Moderator
Jessica S. Ancker, MPH, PhD5

1State University of New York at Buffalo, NY; 2Oregon Health & Science University, Portland, OR; 3University of Victoria, Victoria, BC, Canada; 4The University of Texas Health Science Center, Houston, TX; 5Weill Cornell Medical College, New York, NY

Abstract
Sociotechnical theory has become accepted in the last fifteen years for understanding and evaluating the complexity of design and implementation of health information technology (HIT) in complex health care organizations. The IOM report on information technology and patient safety argues that safety is the product of the sociotechnical system and its constituent parts. The core concept of sociotechnical theory is interdependency. Sociotechnical theory raises a number of methodological questions, since it sits uneasy with accepted research paradigms in biomedicine that focus on measurable clinical outcomes research. The panel seeks to address the value of sociotechnical theory, the underlying research methodologies and to give practical guidance on how studies involving a sociotechnical perspective can be conducted and published.

Keywords:
Biomedical informatics; sociotechnical theory; interpretive research; research metrics; evaluation

Introduction
In its report ‘Health IT and Patient Safety’ the IOM Committee on Patient Safety and Health Information Technology writes that safety is the product of the larger sociotechnical system and emerges from the interaction between different parts of this larger system [1]. The committee defines a sociotechnical system as a collection of hardware and software working in concert within an organization that includes people, processes and technology. Berg, Aarts and van der Lei emphasize the role of the social sciences in understanding sociotechnical systems [2]. From this perspective health care work is a social ‘real life’ phenomenon, which may seem ‘messy’ at first, but is guided by a practical rationality that can be overlooked only at a high price (i.e. systems that compromise patient safety). Therefore, technological innovation is a social process that deeply affects organizations. Finally, through in-depth formative evaluation, system design and implementation can be improved. Klein makes the case that in sociotechnical theory interdependence is the most important concept, and more specifically interdependencies between social and technical aspects of the system (and not social and technical systems) [3].

A quick search of the terms ‘sociotechnical’ and ‘socio-technical’ in the journal title field in PubMed yields 151 publications. Seventeen publications date from before 2000, and 45 papers were published since 2011. This simply means that the term ‘sociotechnical’ is catching on in the health care literature, but does not necessarily imply a consensus on the use of this term.

The increasing perceived relevance of the concept raises a number of issues. First what does the concept mean? It is striking how the concept is used as an adjective: sociotechnical perspective, sociotechnical system, sociotechnical understanding, sociotechnical theory. What is the implication of the adjective for these concepts? What research methods are implied? It is clear that the sociotechnical concept sits uneasily with more accepted methodologies in biomedicine such as the randomized clinical trial. At the one end of the methodological spectrum researchers consider sociotechnical theory as an interpretive theory. It helps them to analyze the descriptions that they obtained through qualitative research methods, such as observing and interviewing. At the other end researchers try to decompose sociotechnical theory into domains, suggesting that each of these domains can be studied
independently. The next question is whether metrics (measurable quantities) can be developed to measure and assess the sociotechnical aspects of a system. These questions are relevant when information technology is embedded into complex health care systems.

This panel is cosponsored by the AMIA Evaluation (Eval) Working Group and People and Organizational Issues (POI) working group.

The panel

This panel seeks to address how researchers utilize sociotechnical thinking in their research projects. Questions that the panelists seek to address are:

• How do they use sociotechnical theory and methods?
• How complete is sociotechnical theory?
• Is sociotechnical understanding of HIT in need of metrics?
• How acceptable is theoretical diversity for health informatics?
• Does sociotechnical theory necessarily imply qualitative research methods?
• How studies involving a sociotechnical perspective be published?

The panelists

For this panel, participants will describe how the theoretical underpinnings of the sociotechnical approach can be applied to real-world, clinical informatics problems and how it has influenced their research methods.

Dr. Aarts’s studies of clinical decision support in electronic medication prescribing lead him to a number of questions about the difficulty of implementing CPOE in different settings. Basically, electronic prescribing based on evidence-based medicine collides with social values that prescribing represent. Understanding these values using sociotechnical theory is crucial to address the design and implementation of electronic prescribing systems [4]. In Dr. Aarts’s perspective sociotechnical theory is a qualitative interpretive theory.

Dr. Kushniruk will describe work in development of an integrative cognitive sociotechnical approach to the design, deployment and testing of health information systems that he co-developed with Dr. Elizabeth Borycki [5]. The approach seeks to integrate methods and frameworks from the study of human factors and cognitive science with socio-technical approaches to analyzing health information system use and impact. To illustrate this multi-faceted approach, examples from the evaluation of electronic health records and medication administration systems will be described. An in-situ approach to conducting clinical simulations that can identify both user and workflow issues that arise in complex health care settings will be described. Implications of the approach for the deployment of systems with increased system-organizational fit and decreased potential for inadvertently introducing technology-induced error will be discussed.

Dr. Sittig will describe an 8-dimension socio-technical model that he co-developed with Hardeep Singh. Briefly, this model and the resulting interactions between the various dimensions of the model has helped them explore and understand clinical informatics problems such as: management of abnormal laboratory test results, development of comparative effectiveness research platforms, and use of a service-oriented architecture approach to the provision of externally managed clinical decision support interventions [6].

Dr. Ash will discuss how this same 8-dimension model has been used in two different ways to expand our knowledge about the safety of electronic health record (EHR) systems. First, she will describe using the model as an organizing framework for synthesizing and interpreting literature published prior to 2010 about EHR safety to produce a white paper which provided the background information for the aforementioned Institute of Medicine study. Next, she will describe how the model informed the research behind the principles and practices recommended in the Safety Assurance Factors for EHR Safety (SAFER) guides developed for the Office of the National Coordinator for Health Information Technology [7].

Following the panelist’s presentations, the moderator will start the discussion by asking a few questions to help the audience understand the similarities and differences among the various approaches. We will then open the floor for a free-ranging discussion of all aspects of the sociotechnical approach.
Learning objectives

After participating in the session, the attendant should be able to:

• Describe how sociotechnical theory can be used in biomedical informatics research.
• Identify in which research project it is appropriate to use sociotechnical theory.
• Understand how to publish studies using sociotechnical theory in biomedical informatics journals.
• Articulate a sociotechnical research agenda for biomedical informatics related to HIT design and implementation.

Conflict of interest

The participants have no conflict of interest to declare.

References

Recent Advances in Computational Drug Repositioning

Atul Butte, MD, PhD¹, Nigam Shah, MBBS, PhD², Nicholas Tatonetti, PhD³, Hua Xu, PhD⁴, Ping Zhang, PhD⁵

¹University of California San Francisco, San Francisco, CA
²Stanford University, Stanford, CA
³Columbia University, New York City, NY
⁴The University of Texas Health Science Center at Houston, Houston, TX
⁵IBM T.J. Watson Research Center, Yorktown Heights, NY

Participants

- Organizer and Moderator: Ping Zhang, PhD, Healthcare Analytics Research, IBM T.J. Watson Research Center
- Panelist: Atul Butte, MD, PhD, Institute of Computational Health Sciences (ICHS), The University of California San Francisco (UCSF)
- Panelist: Nigam Shah, MBBS, PhD, Center for Biomedical Informatics Research, Stanford University
- Panelist: Nicholas Tatonetti, PhD, Departments of Biomedical Informatics, Columbia University
- Panelist: Hua Xu, PhD, School of Biomedical Informatics, The University of Texas Health Science Center at Houston (UTHealth)

Abstract

Computational drug repositioning is a promising and efficient tool for discovering new uses from existing drugs and holds the great potential for precision medicine in the age of big data. The explosive growth of large-scale genomic and phenotypic data, as well data of small molecular compounds with granted regulatory approval, is enabling new developments for computational repositioning. To achieve the shortest path towards new drug indications, advanced data processing and analysis strategies are critical for making sense of these heterogeneous molecular measurements. Despite the progress simulated by big data analytics, there is clearly room for technical improvement with regard to computational drug reppositioning methods. Furthermore, to materialize the true potential and impact of these methods, much work is needed to show that they can be successfully adopted into practical applications. In this panel, participants will summarize the recent advances in computational drug repositioning and identify challenges and opportunities. Panel participants will synthesize their perspectives on these key issues and likely future developments in this area, explore a diverse set of topics, and engage in thoughtful discussion with the audience.

Learning Objectives

After participating in this session, the learner should be better able to:

- Recognize the benefits of applying computational methods for drug repositioning hypothesis generation
- Compare genomic, phenotypic, clinical, epidemiological data, and their contributions to computational methods
- Understand recent advances, challenges, and opportunities of computational drug repositioning
- Evaluate computational repositioning methods from a real-world perspective
- Validate drug repositioning hypotheses in a systematic manner

Introduction

The inefficiency of pharmaceutical drug development with high expenditure but low productivity has been widely discussed. Drug repositioning, the process of finding additional indications (i.e., diseases) for existing drugs, presents a promising avenue for identifying better and safer treatments without the full cost or time required for de novo drug development. Candidates for repositioning are usually either market drugs or drugs that have been
discontinued in clinical trials for reasons other than safety concerns. Because the safety profiles of these drugs are known, clinical trials for alternative indications are cheaper, potentially faster and carry less risk than de novo drug development. Any newly identified indications can be quickly evaluated from phase II clinical trials. Drug repositioning can reduce drug discovery and development time from 10-17 years to potentially 3-12 years. Therefore, it is not surprising that in recent years, new indications, new formulations, and new combinations of previously marketed products accounted for more than 30% of the new medicines that reach their first markets.

Drug repositioning has drawn widespread attention from the pharmaceutical industry, government agencies, and academic institutes. For example, The United States National Center for Advancing Translational Sciences (NCATS) is partnering with pharmaceutical companies to offer funding for repositioning. US Food and Drug Administration (FDA) is also enabling drug repositioning, with the creation of several public databases specifically for computational drug repositioning. All of these efforts significantly promoted drug repositioning research.

**Urgency, Timeliness and Importance of the Panel**

Big data analytics for both drugs and diseases provide an unprecedented opportunity to uncover novel statistical associations between drugs and diseases in a scalable manner. With the attention on translational research in recent years, a new set of computational methods are being developed which examine drug-disease associations and drug off-target effects through system and network approaches. These new approaches take advantage of the unprecedented large-scale high-throughput measurements, such as drug chemical structures and screens, side effect profiles, transcriptional responses after drug treatment, genome wide association studies, and combined knowledge. More importantly there are increasing reports of these findings being validated in experimental models and clinical observations, thus clarifying the value proposition for computational drug discovery. As a result, now is an exciting time for computational scientists to gain evidence for reusing an existing drug for a different use or generate testable hypotheses for further screening.

Despite the progress, there is clearly room for technical improvement with regard to computational drug repositioning approaches. Furthermore, to materialize the true potential and impact of these methods, much work is needed to show that they can be successfully adopted into practical applications. Current challenges include: (1) the transformation of theoretical computational models into practical use if far from straightforward, due to some inevitable factors like missing data, data bias, and technical limitations of computational methods; (2) the lack of structured gold standard (other than cross-validation) for drug repositioning makes it hard to compare and evaluate the performance of computational methods; (3) the standard quantitative measurements in machine learning (e.g., AUC, AUPR, F-score) have their limitations to evaluate computational drug repositioning methods. For example, in real-world drug discovery, a desired method should provide relatively high confidence that the first few indications that are predicted for a drug contain at least one that will be validated in clinical trials and make a positive impact on patients (otherwise the project will be terminated); (4) the lack of standard models for repositioning hypotheses validation (i.e., if one comes up a hypotheses of drug A could potentially treat disease B, how can she/he find the best model to validate this hypothesis) hinders the impact of computational methods in the real-world settings. Therefore, it is very urgent to bring the challenges and opportunities to the table. The aim of our panel session is to provide a forum to bring together the research community for a serious examination of these important issues and to stimulate innovative work for years to come.

This panel gathers five researchers who actively design, develop, and/or validate computational drug repositioning methods. The five panel participants have diverse backgrounds (e.g., academia and industry, MBBS/MD and PhD, big data and biomedicine) and have their own perspectives on these challenges and opportunities. All participants will share experiences in conducting their research in computational drug repositioning and discuss the implications of their studies. These individual stories will then serve as the material for an open discussion with the audience. In sum, this is an important panel to have at AMIA in order to engage the community in discussing critical questions about advancing computational drug repositioning, which shortens the drug development timeline and gets new treatments to patients faster.

**Brief Description of Panelists and Presentations**

The panel session will begin with four 10-minute presentations of panelists (each followed by 5 minutes of questions) detailing experiences, approaches and outcomes of their exemplar studies. Then we will open the floor to discussion among panelists and with the audience in the remaining 30 minutes.

**Atul Butte, MD, PhD** is the new Director of the new Institute of Computational Health Sciences (ICHS) at the University of California, San Francisco, and a Professor of Pediatrics. Dr. Butte will present his experience in
translating trillions points of data into new uses for drugs. Dr. Butte's lab builds and applies tools that convert trillions points of molecular, clinical, and epidemiological data -- measured by researchers and clinicians over the past decade and now commonly termed “big data” -- into diagnostics, therapeutics, and new insights into disease. He will highlight how publicly available molecular measurements to find new uses for drugs including drug repositioning for inflammatory bowel disease, as well as discovering new treatable inflammatory mechanisms of disease in type 2 diabetes, and how the next generation of biotech companies might even start in your garage.

Nigam Shah, MBBS, PhD is Assistant Professor of Medicine (Biomedical Informatics) at Stanford University, Assistant Director of the Center for Biomedical Informatics Research, and a core member of the Biomedical Informatics Graduate Program. Dr. Shah will present his recent work on automated detection of off-label drug use, which is a form of drug repositioning because such uses have not been evaluated for safety and efficacy. He will describe a data-mining approach for systematically identifying off-label usages using features derived from free text clinical notes and features extracted from two databases on known usage (Medi-Span and DrugBank).

Nicholas Tatonetti, PhD is Assistant Professor of Biomedical Informatics in the Departments of Biomedical Informatics, Systems Biology, and Medicine and is Director of Clinical Informatics at the Herbert Irving Comprehensive Cancer Center at Columbia University. Dr. Tatonetti will speak about network medicine, which integrates experimental data on gene, protein, and metabolic interactions with clinical knowledge of disease and pharmacology in order to extend the understanding of diseases and their treatments. He will highlight a case study that uses network analysis and graph theory for understanding drug safety and efficacy, which recently published in Clinical Pharmacology & Therapeutics.

Hua Xu, PhD is Robert H. Graham Associate Professor at the School of Biomedical Informatics in The University of Texas Health Science Center at Houston (UTHealth). He directs the Center for Computational Biomedicine at UTHealth. Dr. Xu will present his recent work to assess the feasibility of using electronic health records (EHRs) and automated informatics methods to efficiently validate a recent drug repurposing association of metformin reducing cancer mortality. This study serves as a model for robust and inexpensive validation studies for drug repurposing signals using EHR data.

Ping Zhang, PhD is Research Staff Member in IBM T. J. Watson Research Center. Dr. Zhang will moderate the panel. During the panel discussion, Dr. Zhang will discuss how to evaluate and validate drug repositioning methods in real-world settings.

Discussion Questions
A list of discussion questions to enhance audience participation:

1. How will computational methods accelerate drug repositioning (or even drug development in general) efforts?
2. Lots of data are used for drug repositioning, e.g., genomic, phenotypic, clinical, epidemiological data. What are the pros and cons for each type of data?
3. How to integrate disparate sources of data for drug repositioning purpose? Is there any prior weight for each data source?
4. What is the success rate of drug repositioning methods in a real-world setting? How much improvement after using computational methods? How to transfer computational models into practical uses?
5. How to design an experiment to compare multiple computational methods for a repositioning purpose? How to prepare/define gold standard? Which evaluation metrics should choose? Do current metrics meet the needs from pharmaceutical industry?
6. How to validate repositioning candidates? How to identify a best model for a specific hypotheses validation? What are the pros and cons for literature/experimental/clinical validations? How to leverage real-world data to validate repositioning candidates?

Participation Statement
All proposed participants have agreed to take part on the panel.
Health Data from Wearables:
Clinical Implications of the Quantified Self

Mobile Health Track

Panel Organizer and Participants
Emil Chiauzzi, Ph.D. (organizer and presenter)
Research Director
PatientsLikeMe
Cambridge, MA

Kevin Patrick, M.D., M.S. (moderator and presenter)
Professor of Family Medicine & Public Health
Director, Center for Wireless and Population Health Systems, The Qualcomm Institute
University of California, San Diego
San Diego, CA

Cinnamon Bloss, Ph.D. (presenter)
Assistant Professor, Departments of Psychiatry and Family Medicine & Public Health, Division of Health Policy
University of California, San Diego
San Diego, CA

Ernesto Ramirez (presenter)
Program Director
Quantified Self Labs
San Diego, CA

All four original panel members have agreed to participate in the panel. We will add a patient to the panel prior to the conference.

Abstract
The popularity of health self-tracking through the use of consumer health devices has primarily been confined to fitness and wellness contexts, but there has been increasing interest in leveraging this health data in clinical decision-making and care delivery. Because personal tracking devices provide a glimpse into the everyday behavior of patients, there is potentially great value in integrating this data with traditional medical data. Such integration may be particularly useful with patients experiencing chronic diseases, as their conditions require daily self-management that include activity goals, sleep monitoring, an understanding of calorie expenditure, and tracking of medication adherence. However, there are many questions that are unresolved. There is a need for clarity about how such data can be utilized in clinical encounters. There are numerous devices available to consumers and they vary greatly in features and data quality. The patient perspective has not been fully leveraged, as devices generally target a fitness market. Users of tracking devices may not be fully cognizant of issues around privacy, data sharing, and data control. Most importantly, does the use of self-tracking devices affect disease outcomes?

Description of Panel
This panel will examine personal health data from several perspectives; (1) the use of data by clinicians, (2) the validity of wearable data and the potential capabilities of wearables going forward, (3) patient perspectives on the use of wearables in disease self-management, and (4) ethical and privacy issues in the use and transmission of this data.

Kevin Patrick (Moderator)
Personal health data promises to greatly enhance how clinicians manage their patients’ conditions. Instead of relying on periodic self-reports of behaviors or symptoms, data accumulated on a 24x7 basis from wearable and wireless technologies can be summarized and reviewed. Analysis of patterns in these data can be used to infer trajectories of illness and health and aid in developing just-in-time adaptive interventions that might prevent deterioration in health status. However, many challenges exist with respect to factoring personal health data into the workflow of busy
clinicians. This presentation will provide an overview of these issues.

**Ernesto Ramirez**
Self-tracking data are becoming a part of the normal course of everyday life for millions of individuals around the world. What are some of the ways they are using that data to understand themselves and their health? We will highlight the unique and innovative methods and insights being discovered and discussed at Quantified Self meetings and events worldwide. Particular focus will be given to personal data access and its role in exposing up new knowledge generation methods for personal and public benefit. Additionally, we’ll discuss issues with data validity and reliability regarding consumer sensing technologies.

**Emil Chiauzzi**
The use of wearables has allowed patients with chronic diseases to take an active role in self-management. This presentation will review insights from a series of pilot studies conducted with members of PatientsLikeMe, an online patient research network. This presentation will review the ways in which patients have integrated wearable use into their treatment, as well as their perspectives on behavior change methods that can be integrated with wearable use. The importance of engagement strategies will also be highlighted.

**Cinnamon Bloss**
Personal health data that circulate outside the mainstream of traditional health care have the potential to be used in ways that could improve health, in part, by empowering individuals to take greater control over their own health. In sharp contrast to the idea of individual empowerment around health, however, is a personal health data ecosystem in which individuals may feel they have little control over the flow of their personal information and thus privacy. This presentation will discuss issues related to personal conceptions of privacy and health data ownership. Preliminary data on individuals’ perceptions and expectations of privacy in the context of different personal health data technologies will be presented.

**Patient Presenter (TBD)**
The use of wearables to manage chronic disease offers a different set of challenges than using these devices for personal fitness. The fluctuations in symptoms, physical limitations, effects of pain, and other factors change the ways in which a wearables device is used on a daily basis. Many patients find unique ways to integrate these devices (and their data) into self-management. The limitations and advantages of these devices in a chronic disease context will be discussed.

**Explanation of Topic**
The explosion of consumer health devices, punctuated by the publicity surrounding devices such as the Apple Watch, makes this a very timely topic. There is a growing research and clinical literature on the application of wearables beyond fitness contexts to clinical contexts. At the same time there are a variety of technological and practice barriers that need to be addressed. This panel seems best geared for audience that is composed of clinicians, researchers, and informaticists who are curious about how this data can be maximized in their work.

**Discussion Questions**
What is the role of wearable data in medical care?
How can wearable data best be integrated into the clinical workflow?
How do patients perceive the usefulness of devices in their disease self-management?
What clinical assessment needs can personal health data best address?
What is the potential for wearable use to affect health outcomes?
What adaptations must patients make to use wearables effectively?
What are key privacy concerns in the use of wearable data?
The Informatics Workforce for Population Health: Challenges, Initiatives and the Path Forward

Brian E. Dixon, MPA, PhD1,2, Marty LaVenture, PhD, MPH3, Bill Brand, MPH4, Arthur Davidson, MD, MSPH5, Shaun J. Grannis, MD, MS2,6
1IU Richard M. Fairbanks School of Public Health, Indianapolis, IN; 2Regenstrief Institute Center for Biomedical Informatics, Indianapolis, IN; 3Minnesota Department of Health, St. Paul, MN; 4Public Health Informatics Institute, Decatur, GA; 5Denver Public Health, Denver, CO; 6Indiana University School of Medicine, Indianapolis, IN

Abstract
Public health informatics (PHI) makes significant contributions to detecting, monitoring and improving population health. Although PHI professionals exist in local, state and federal public health agencies, we know little about them or their needs. Yet there exist several disparate education and training programs in PHI that seek to prepare the public health workforce for designing, implementing, and using PHI systems. This panel presents diverse views on the PHI workforce and its needs from local, state, federal and international levels. Current data on the characteristics and information needs of the PHI workforce will be presented. Then panelists will discuss the various initiatives and programs that seek to educate or train the PHI workforce. The panel will raise questions and discuss opportunities about how to better organize information about training opportunities and the role that AMIA should play in supporting the existing and future PHI workforce.

Introduction and Background
Public health informatics (PHI) is the systematic application of information and computer science as well as information systems to public health practice, research, and learning (1). Although public health practitioners have long utilized information technologies to perform their jobs, the growth of PHI as a discipline within both public health and the broader field of informatics accelerated at the start of the twenty-first century.

During the first decade, PHI activities were characterized by a focus on the core public health function of monitoring populations: early detection of bioterrorism, such as the Anthrax attacks in the U.S. (2) and the Tokyo subway attacks (3), as well as global health threats such as SARS (4) and the H1N1 pandemic (5). While the threat of a large-scale epidemic has not diminished in recent years, as evidenced in 2014 by MERS (6, 7) and Ebola (8), changes in national policies and funding priorities have steered PHI in new directions (9). Today PHI not only supports core public health functions, but PHI contributes to the following activities in support of population health and strengthening the public health infrastructure: 1) implementation of electronic health record (EHR) systems and health information exchange (HIE) to enable successful achievement of “meaningful use” criteria such as electronic reporting of notifiable diseases; 2) measurement of a wider array of health indicators, including social determinants of health through “big data” analysis of multiple community data sources; and 3) development, implementation, and assessment of patient-centered technologies aimed at supporting health and well-being in the changing landscape of health care delivery locations such as the patient’s home, community organizations, and pharmacy.

Systematic downsizing of public health departments across the United States over the past decade (10, 11) has created the need for an estimated 250,000 workers by 2020 to maintain current capacity for public health services (12). Today, given the growing demand for a PHI capable workforce, key stakeholders recognize PHI as an important core within modern public health practice (13-15). Yet only a fraction of the public health workforce is estimated to have received informatics training or specialization in PHI, and we know little about the characteristics (e.g., race, age) or information needs (e.g., competency skill level) of current PHI workers.

Panel Description
This panel will review the state of the existing PHI workforce, discuss the challenges in cultivating a robust workforce, and brief the audience on existing initiatives aimed at preparing the next generation of PHI leaders. Each panelist brings a unique perspective and contribution to the common theme of strengthening the PHI workforce. Together they will provide views from research, education and operations/practice. Attendees will leave with a sense of where things stand as well as directions the field is headed to mitigate the need for a highly trained PHI workforce. Given that AMIA strives to be the professional home for PHI, it is well suited for the annual conference.
Current State of the PHI Workforce (Dixon)

The panel will begin with a brief review of the existing state of the PHI workforce by Brian E. Dixon, PhD. Dr. Dixon is a research scientist at the Regenstrief Institute and faculty at the Indiana University Fairbanks School of Public Health. He further serves on the Executive Committee of the PHI Working Group within AMIA.

While organizations such as the National Association of City and County Health Officials (NACCHO) and the Association of State and Territorial Health Officers (ASTHO) gather data on the public health workforce, they have not comprehensively examined the PHI segment. Currently Dr. Dixon is leading a team of researchers to analyze the Public Health Workforce Information Needs Survey (PH WINS) conducted by ASTHO in late 2014. They survey includes responses from over 10,000 public health workers in state health agencies. Just 1.3% of the respondents indicated they work in a PHI role within their agency. In his presentation, Dr. Dixon will describe the survey and characterize the population of PHI respondents. He will further describe their information needs, and he will compare/contrast the ASTHO results with data from other workforce surveys. Relevant data from prior studies will also be presented as needed. The goal is to provide attendees with up-to-date information on the workforce and its unique information needs.

Existing Education and Training Efforts: View from the National Level (Brand)

Bill Brand is the Director of Informatics Science at the Public Health Informatics Institute (PHII) in Decatur Georgia. Mr. Brand will review several examples of PHI training programs, both for emerging PHI professionals and for public health practitioners who realize that PHI knowledge and skills are critical to being effective in their jobs. The review will include ONC certificate programs, various CDC fellowships, the PHII Informatics Academy, and the PHII Immunization Registry Workforce Development project. He will describe these projects, their aims, and how they contribute to the overall workforce development.

Initiatives and Views from the State Level (LaVenture)

Dr. Marty LaVenture heads the Office of Health Information Technology at the Minnesota Department of Health and leads the Minnesota e-Health Initiative. Dr. LaVenture will discuss the PHI workforce from a state health agency perspective. Specifically, he will discuss the status an ASTHO state public health informatics directors initiative to promote informatics roles and career series for public health informaticians. Furthermore, Dr. LaVenture currently represents the public health community on AMIA’s Advanced Interprofessional Informatics Certification (AIIC) Program steering committee, so he will provide a brief update on that initiative. In addition, he will discuss achieving balance between academic learning and immediate PHI training needs. He will further discuss opportunities for graduate student practicums and other project-based work to contribute to the overall development of the PHI workforce as part of achieving an informatics savvy organization. (16)

View from the Local Level (Davidson)

Art Davidson, MD is Director of Informatics, Epidemiology and Preparedness at Denver Public Health, Denver Health. Dr. Davidson will describe the informatics workforce needs and challenges within the Denver Metro region and nationally based on efforts and concerns from National Association of County and City Health Officials (NACCHO) colleagues. Dr. Davidson serves as co-chair for the interim executive committee of the ASTHO-housed, Public Health Community Platform – a multi-year CDC-funded initiative to support a community-driven, valued service. Particular local public health agency challenges to be addressed will include ability to develop local applications despite a constrained informatics workforce, key skills for application development, and strategic efforts around regional or national platforms. Issues of organizational role, recruitment, promotion and retention of skilled informaticians within a local health department will be discussed.

A View of the PHI Workforce from the International Level (Grannis)

Outside the U.S., PHI is commonly performed in the context of a Ministry of Health, which not only supports governmental public health functionality, but also serves as the administrator for the country’s health system. In this presentation, Dr. Shaun Grannis, Associate Professor of Family Medicine at the Indiana University School of Medicine and Associate Director of the Regenstrief Institute’s Center for Biomedical Informatics, will draw parallels between PHI and clinical informatics in the context of global health informatics activities. The capacity development needs (and efforts) for global health informatics will be compared to the activities of Ministries to illustrate how such efforts are complementary to emerging U.S. strategies for local, state, and federal governmental health agencies.
Discussion Questions

1. How should AMIA best support the growth and maintenance of the PHI workforce?
2. What is the role for AMIA in contributing to the content of PHI training programs?
3. How can AMIA support a better coordinated approach to collecting and disseminating information about PHI training programs and opportunities?
4. What role should AMIA play in certifying core competencies for those who choose to specialize in PHI?

Panel Organizer Statement

All participants have agreed to take part in the panel and discuss the topics as outlined above.

References

Looking Back and Moving Forward: A Review of Public and Global Health Informatics Literature and Events

Brian E. Dixon, MPA, PhD\textsuperscript{a,b,c}, Jamie Pina, PhD, MSPH\textsuperscript{d,e}, Janise Richards, PhD, MPH, MS\textsuperscript{f}, Hadi Kharrazi, MD\textsuperscript{g}, Anne Turner, PhD, MLIS, MPH\textsuperscript{h}

\textsuperscript{a} Indiana University Fairbanks School of Public Health, Indianapolis, IN
\textsuperscript{b} Regenstrief Institute, Center for Biomedical Informatics, Indianapolis, IN
\textsuperscript{c} Center for Health Information and Communication, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416, Indianapolis, IN
\textsuperscript{d} RTI International, Waltham, MA
\textsuperscript{e} Emory University, Atlanta, GA
\textsuperscript{f} Centers for Disease Control and Prevention, Center for Global Health, Atlanta GA
\textsuperscript{g} Department of Health Policy and Management, Johns Hopkins School of Public Health, Baltimore, MD
\textsuperscript{h} Department of Biomedical Informatics and Medical Education, University of Washington School of Public Health, Seattle, WA

Abstract

Over the past decade, the disciplines of public health and global health informatics have rapidly expanded and continue to evolve within the field of biomedical informatics. Increased attention and activity by the U.S. Centers for Disease Control and Prevention as well as many health ministries, the World Health Organization, and non-governmental organizations are generating new knowledge and lessons regarding the development, implementation, and use of information systems in health care delivery around the globe. Thus, a growing body of literature now contains important insights and lessons from international informatics activities, stimulating the need to synthesize the knowledge for the field. In this panel, a review of recent literature in the areas of public health and global health informatics will be presented. Key articles revealing trends, methods, and lessons will be summarized to bring attendees up-to-date on the use of informatics to improve population health as well as care in low resource settings.

Introduction

At the AMIA 2013 and 2014 Annual Symposia, Drs. Brian Dixon and Jamie Pina presented well-attended panel presentations focused on summarizing recent literature and key activities in Public Health Informatics (PHI) and Global Health Informatics (GHI). We seek to present a similar panel at the 2015 symposium. Our goal is to provide a concise overview of landmark papers in PHI and GHI for the year prior to the 2015 symposium using the methods described below and to highlight key activities that have impacted PHI and GHI practice and research in the past 12 months. Furthermore, like the last two years, we will provide the output of our efforts as working documents available online enabling attendees, as well as those who cannot attend, to leverage our work throughout the year.

Overview

The fields of public and global health informatics are expanding. In 2012, the PHI Fellowship Program administered by the U.S. Centers for Disease Control and Prevention (CDC) became a Department of Labor Registered Apprenticeship. PHI was also one of the six sub-disciplines of informatics specified in the ONC certificate programs funded between 2010 and 2013. As a result, publications in PHI have dramatically increased. This is evidenced by an increase in \textit{Journal of the American Medical Informatics Association} (\textit{JAMIA}) PHI articles, as well as the emergence of a journal dedicated completed to the study and practice of PHI, the \textit{Online Journal of Public Health Informatics} (www.ojphi.org). Similar increases in programs and publications have been observed in the field of GHI. For example, Fogerty International fellowships now bring scholars from low-income counties to the U.S. for biomedical informatics training. The CDC also promotes a Global Public Health Informatics Program (http://www.cdc.gov/globalhealth/dphswd/gphip/what/objectives.htm) to build capacity and sustainability of health information systems world-wide. The impact factor and quality of publications in the \textit{International Journal of
Medical Informatics (IJMI) has also increased, leading to a significantly larger number of important papers published by non-U.S. scholars in informatics (1).

Given recent and continuing advances in PHI and GHI as well as the organizational resources committed by AMIA to the PHI and GHI Working Groups as well as its International Affairs Committee, an annual review of landmark publications in these sub-disciplines within the field of biomedical informatics is appropriate for the AMIA Annual Symposium. Two panelists (BED and JP) will each present a concise bibliography of recent landmark publications in either PHI or GHI. The session will be divided equally between PHI and GHI. A summary examining the convergence, synergies, priorities and possible future cooperation will be presented by the third panelist (JR). Time will be reserved at the end for discussion, question, and comments with the audience.

Methods

The methods used to select the literature in each part of the session are adapted from the work of the previous AMIA Year in Review sessions. Our search methodology to identify potentially high impact literature in PHI and GHI is outlined below. The methods will be executed using MEDLINE and selected resources not currently indexed such as the conference proceedings of the International Society for Disease Surveillance. Recommendations will also be solicited from the AMIA PHI and GHI working groups as well as International Affairs Committee members.

Candidate articles will be reviewed on a quarterly basis by a committee of experts in PHI and GHI, including Drs. Dixon, Turner, Kharrazi, Pina, and Richards, the Review’s core team members. Additional input from the AMIA PHI and GHI working groups will be solicited. Selected articles from the list of candidates will be determined through consensus discussions on quarterly conference calls. Final lists of selected articles will be made available online for reference following the symposium.

Public Health Informatics

In 2013 and 2014 to identify articles in PHI, we used the following query (customized for MEDLINE):

“Public Health Informatics”[mh] OR (“exp Public Health”[mh] and “exp Informatics”[mh]) OR (“public health”[mp] and “informatics”[mp])

This query in 2013 returned 392 results, after specific inclusion and exclusion criteria were applied 65 articles were fully reviewed. In 2014 the query returned 160 results, with 20 articles meeting the criteria for full review. While this query yielded a reasonable set for the committee to review and discuss within a short timeframe, the group feels the query is not sensitive enough to capture “information system” articles that fail to use the term “informatics.” Therefore we propose using the following query for 2015:

“Public Health Informatics”[mh] OR (“exp Public Health”[mh] and “exp Informatics”[mh]) OR (“public health”[mp] and “informatics”[mp]) OR (“public health”[mp] and “information system”[mp]) OR (“public health”[mp] and “information technology”[mp]) OR (“public health”[mp] and “electronic health record”[mh])

In the 2015 presentation, Dr. Dixon will present the final selection of papers from the set of PHI candidate papers. Papers may be further categorized into themes, such as disease surveillance, population health, public health decision support, or chronic disease management depending on the list of candidate papers and the perspectives of the review committee. When reviewing the papers, Dr. Dixon will briefly describe their aim, methods, and key findings. Paper results will also be discussed in the context of larger PHI developments, such as the release (or study) of a specific system (e.g., Biosense 2.0) or change in public health infrastructure (e.g., re-organization of the CDC’s division responsible for PHI). Context will help audience members connect specific study findings with larger trends within the biomedical informatics field of study and practice.

Global Health Informatics

In 2013 to identify articles in GHI, we used the following query (customized for MEDLINE):

(“exp Informatics”[mh] OR “exp Telemedicine”[mh] OR “information system”[mp]) AND (“Developing Countries”[mh] OR global OR ministry OR “low resource” or “resource-limited”[mp])

This query with a search range of 24 months returned 442 results. During the review process the GHI reviewers thought this query missed many known articles. To overcome these limitations, we worked closely with an information specialist at the CDC Library to develop this more sensitive query for our 2014 MEDLINE search:
This query with a search range of 12 months (11/01/2013-10/31/2014) returned 819 results. After a title and abstract review based on seven inclusion criteria, 199 articles were included in the final analysis. For 2015, we plan to further refine this search string to maintain sensitivity while increasing specificity to reduce non-GHI articles for final analysis.

In the 2015 presentation, Dr. Pina will present the final selection of papers from the set of GHI candidate papers. Papers may be further categorized into themes, such as mHealth, infrastructure, evaluation, capacity development, system implementation, etc. depending on candidate papers located and the review committee’s assessments. The reviewers will briefly describe the paper’s aim, methods, and key findings. Findings from the trends in the papers reviewed will also be discussed in context of global health or global informatics policy related activities, such as World Health Organization (WHO) guidelines, the US Congress reauthorization of the President’s Emergency Plan For AIDs Relief (PEPFAR) and the increasing number of individual country eHealth Strategy Policy documents.

Following the PHI and GHI presentations, Dr. Richards will provide a summary to examine findings to provide context for possible PHI and GHI priority activities for the coming year. The session will end with discussion. Our experience from last two years of presenting our findings has demonstrated that these related but unique sub-fields within biomedical informatics generate many interesting questions and comments from the audience and facilitate a dialogue between AMIA’s international and domestic membership.

Discussion Questions

We will draw from the following discussion questions to engage audience participation:

- We found several interesting trends in knowledge generation within PHI & GHI over the past year. Which trend(s) were surprising to you?
- Projecting on what we have found, what do you think will be the biggest trend for 2016?
- What trends do you know of in GHI and PHI for the past year that may not have made it into the literature, yet?
- What initiatives, organizations and/or funding may exist to drive improvements in PHI & GHI research methods as well as dissemination and translation into practice?
- How can AMIA, especially the PHI & GHI working groups, better support the generation of new knowledge for research and practice in our respective disciplines?

Panel Organizer Statement

All participants have agreed to take part in the panel and discuss the topics as outlined above.

References

Building an Open, Global Research Collaboratory: Lessons from the OHDSI Initiative

Jon D. Duke, MD MS¹, George Hripcsak, MD MS², Patrick Ryan PhD³, Nigam Shah MBBS, PhD⁴
¹Regenstrief Institute, Indianapolis, IN; ²Columbia University, New York, NY; ³Janssen Research and Development, Titusville, NJ; ⁴Stanford University, Palo Alto, CA

Abstract

The Observational Health Data Sciences and Informatics collaborative (OHDSI, pronounced ‘Odyssey’) was formed in 2013 with the goal of creating reliable scientific evidence through large-scale analysis of observational health data from around the world. To advance this goal, OHDSI has had to grow rapidly its scientific, technical, and community infrastructure. OHDSI has now grown to over 120 participants from 10 countries. Across the collaborative, there are over 50 databases covering hundreds of millions of patient lives. OHDSI has conducted multiple international network-based observational research studies using this infrastructure. In this panel, we present initial results and lessons learned from the OHDSI initiative.

Organizer: Jon D. Duke, MD, MS

Background and Panel Description

The Observational Health Data Sciences and Informatics collaborative (OHDSI, pronounced ‘Odyssey’) was formed in 2013 with the goal of creating reliable scientific evidence through large-scale analysis of observational health data from around the world. To advance this goal, OHDSI has had to grow rapidly its scientific, technical, and community infrastructure. OHDSI has now grown to over 120 participants from 10 countries. Across the collaborative, there are over 50 databases covering hundreds of millions of patient lives that have been transformed into a common data model. OHDSI has conducted multiple international network-based observational research studies using this infrastructure. In this panel, we present initial results and lessons learned from the OHDSI initiative.

Specifically, we will demonstrate how an international distributed network can be used to generate clinically meaningful results from disparate observational databases, using a systematic evaluation of the heterogeneity in patient treatment progression and its relation to current clinical guidelines as a motivating example. Areas of discussion will include technical challenges, such as accommodating diverse data sources, as well as social challenges, such as coordinating activities amongst diverse stakeholders with differing objectives including those in healthcare, academia, government, and industry. We will also explore the many rewards of a large international collaborative, including the sharing of expertise and the ability to execute multi-site research studies in a streamlined fashion.

By the end of this panel, attendees will gain an understanding of the issues and approaches to collaborative observational research.

Panel Importance and Target Audience

The proposed panel demonstrates the potential for international collaborative research and addresses the critical need for both institutional and individual stakeholders to leverage the rapid advances in collaborative data science. This panel will provide attendees with insight into the opportunities and challenges of an open research collaborative as well as a clear path to engaging in such initiatives. The target audience includes health informaticians, clinical and health services researchers, health IT leaders (e.g., CMIOs), implementation scientists, epidemiologists, biostatisticians, and data scientists.
Presentations

Vision for Research Studies in an Open Observational Research Network (G. Hripcsak)
The mission of OHDSI is to generate high quality medical evidence through large-scale observational research. Dr. Hripcsak will describe the scope of OHDSI and the vision for the kinds of research that may be supported and the current and planned mechanisms to facilitate such research. He will introduce a proof-of-concept study on treatment pathways to characterize the diversity of initial treatment pathways for three chronic diseases: Hypertension, Type II Diabetes, and Depression. While numerous treatment guidelines exist for chronic conditions, there is a paucity of data on the real-world treatment pathways that patients experience in practice.

Proof of Concept: a Large-Scale Study of Treatment Pathways (N. Shah)
Dr. Shah will describe the methods and results from the treatment pathways study. He will summarize the treatment pathways observed among patients who have at least 12 months of continuous observation and ongoing treatment following drug initiation. He will describe temporal trends, and characterize variations in treatment pathways by population, geography, and data capture process. Figure 1 shows a visualization of results from the hypertension pathways study.

Architecting a Common Data Model for Large-Scale Analytics (P. Ryan)
OHDSI utilizes the OMOP Common Data Model (CDM). First developed in 2008, the CDM was designed to enable standardized analysis of patient-level observational data to produce population-level real-world evidence from across multiple disparate data sources. Over 50 papers have been published using the CDM as a data model. Dr. Ryan will cover the design of the CDM and its standardized vocabularies as well as its evolution from supporting a pharmacovigilance methods research project to becoming a broader framework for collaborative research in healthcare. He will discuss how the architecture benefited the treatment pathways study, and he will discuss the challenges encountered when executing the study, primarily due to residuals differences among sites.

Transforming Data to the OMOP Model: Tools and Strategies for Data Owners (J. Duke)
For organizations wanting to participate in the OHDSI research network, transforming their data into the OMOP CDM is a necessary first step. Converting data from a complex existing source, such as an EHR, requires careful planning and the involvement of several key stakeholders (experts in the local source data, clinicians, developers) as well as support from the OHDSI community. In this presentation, Dr. Duke will discuss strategies for the OMOP ETL process as well as some specific software tools and community resources to aid in data transformation. He will then discuss an OHDSI software package known as ACHILLES (www.ohdsi.org/web/achilles), designed to assist with dataset characterization, quality assessment, and exploration of newly converted CDM data (Figure 2).
Figure 2. A screenshot of the ACHILLES data characterization and quality tool, assessing data density over time.

Discussion Questions

- What are the main challenges associated with distributed research?
- What are the costs and privacy risks of participating in a large network research?
- How do we know when to trust (or not trust) the results of a distributed study?
- What is gained and what is lost in the process of standardizing data to a common data model?
- How can standardized analytics be applied in a distributed network using a common data model?
- What are the domain and technical skillsets needed for transforming datasets from one model to another?
- How can the quality of a transformed data set be evaluated and, where necessary, improved?

Assurance

The organizer Jon D. Duke vouches that all listed participants have agreed to take part on the panel.

References

ClinicalTrials.gov: Adding Value through Informatics

Vojtech Huser, MD, PhD¹, Alexa T. McCray, PhD², Neil R. Smalheiser, MD, PhD³,
Asba Tasneem, PhD⁴, Chunhua Weng, PhD⁵

¹Clinical Center, National Institutes of Health, Bethesda, MD;
²Center for Biomedical Informatics, Harvard Medical School, Boston, MA;
³Department of Psychiatry, University of Illinois at Chicago College of Medicine, Chicago, IL;
⁴Duke Translational Medicine Institute, Duke University, Durham, NC;
⁵Department of Biomedical Informatics, Columbia University, New York, NY.

Abstract

ClinicalTrials.gov is a repository of registered clinical trials maintained by NLM containing detailed descriptions of trial sponsorship, design, and results (when available). ClinicalTrials.gov plays an increasingly important, pivotal role in evidence-based medicine, and serves a diverse audience ranging from clinical researchers, who are designing and conducting new trials and recruiting patients; systematic reviewers, who are summarizing the best available evidence regarding safety and efficacy; bio-entrepreneurs, who are looking for drug repurposing or new therapeutic opportunities; and patients, who may be looking for a suitable clinical trial that might accept them. This panel will present an overview of ClinicalTrials.gov and discuss several ongoing lines of informatics research that are adding value -- for example, using text mining to improve the computability of eligibility criteria, design attributes and outcome results and connect these with patient EHR data; linking a registered trial with the publications arising from that trial; and performing aggregate analyses across trials to extract reusable design knowledge, understand design patterns and trends, and uncover systematic biases. The panelists will also discuss challenges and opportunities for further evolution of ClinicalTrials.gov, particularly in light of emerging trends such as patient-centered clinical trials, or the use of unpublished trial data in meta-analyses.

General description of issues examined and brief description of each presentation

Alexa McCray will present an overview of ClinicalTrials.gov: its history, structure, policies, growth, evolution to include trial results, and current user populations¹,²,³.

Vojtech Huser will discuss studies on the completeness and accuracy of ClinicalTrials.gov data. He will review strategies for linking PubMed publications to the appropriate registered trial, as well as recent initiatives to make unpublished and raw trial data publicly available for analysis. He will also show examples how existing summary results data can be used in pharmacovigilance research⁴,⁵,⁶,⁷,⁸.

Asba Tasneem will describe the creation of an Analysis dataset: a database for Aggregate Analysis of ClinicalTrials.gov (AACT)⁹: its design, available database extract formats, utility for aggregate analysis from this dataset, and updating strategy. She will also present trends and outliers of aggregate analysis using ClinicalTrials.gov dataset¹⁰.

Chunhua Weng will describe a suite of text mining algorithms for ClinicalTrials.gov called EliXR. Then she will introduce COMPACT, a database of discrete clinical trial eligibility features, and VITTA, a scalable method for visual aggregate analysis of clinical trial eligibility features across multiple trials. Leveraging this analytical infrastructure, she will present novel findings about the patterns, trends, and biases of clinical trial designs and their implications for various clinical trial stakeholders¹¹,¹²,¹³,¹⁴.

Why is this topic timely? Who is the anticipated audience?

ClinicalTrials.gov is a major national resource that is rapidly growing both in size and scope. Recent changes in Federal policy have expanded the mandates for registering trials and reporting findings. The relatively recent establishment of the Clinical and Translational Science Awards (CTSA) Program and the National Center for Advancing Translational Sciences (NCATS) are two indicators of the fundamental importance of translational medicine to re-engineering the US health care enterprise, and management of clinical trial information lies at the heart of this effort. Many leading
informatics research groups have been using ClinicalTrials.gov as a test bed for advancing the state of the art of information extraction, data mining and visualization to assist in decision making.

For all of these reasons, our panel topic is timely and should draw a wide cross-section of AMIA attendees, including both academic researchers and hospital-based practitioners involved in Clinical Research Informatics. As well, the audience should include researchers, practitioners, health policy makers, and patient advocates, who are not currently involved in handling clinical trial data, but who are looking for opportunities to apply informatics techniques (e.g. text mining, EHR design, and data warehouses) to improve the generalizability of clinical trial designs and the efficiency of clinical trial conduct, and to make the design rationale more transparent to the public.

**List of discussion questions to enhance audience participation**

An increasing importance of ClinicalTrials.gov for clinical research informatics researchers is evidenced by the recent NIH policy update in November 2014 that aims to promote transparency for all NIH-funded clinical trials, whether or not they are subject to mandatory registration or results reporting. What other new Federal policies and national initiatives are on the horizon, which are likely to impact on ClinicalTrials.gov?

What impact is the Notice of Proposed Rulemaking (NPRM) for Clinical Trials Registration and Results Submission likely to have on greater sharing of clinical trial information and how can the informatics community contribute to maximizing that impact?

Published papers arising from a clinical trial often report different outcome measures than what was specified in the original registered trial. Similarly, adverse effects are often under-reported in the published paper relative to the results available in ClinicalTrials.gov. How can ClinicalTrials.gov be mined so as to more readily compare the data associated with a registered trial with data that is eventually published?

How are patients utilizing ClinicalTrials.gov? Are there challenges or opportunities for informatics to improve their experience finding relevant trials? What roles can patients play to improve clinical trial designs and conduct?

How can we engage different stakeholders to improve clinical trial designs and conduct?

**Statement from the Panel Organizer**

I attest that all participants have agreed to take part on the panel, and have contributed to the writing of the proposal.

**References**

Clinical Decision Support: How to Apply Standards to Deliver Knowledge-Driven Interventions

Robert A. Jenders, MD, MS, FACP, FACMI1,2, Guilherme Del Fiol, MD, PhD3
Kensaku Kawamoto, MD, PhD, MHS3, Howard R. Strasberg, MD, MS, FACMI4

1Charles Drew University, Los Angeles, CA
2University of California, Los Angeles, CA
3University of Utah, Salt Lake City, UT
4Wolters Kluwer Health, Sunnyvale, CA

Abstract
Clinical decision support (CDS) is receiving increasing attention as a way to help improve clinical practice and health behaviors. The use of health information technology (HIT) standards for encoding data, representing knowledge and delivering knowledge-based interventions can help facilitate implementation of CDS. However, many standards from numerous standards development organizations (SDOs) exist that are variously incorporated into vendor software, and consensus on the use of these standards is lacking. Moreover, newly proposed standards elaborated in the past year in CDS and related areas such as clinical quality measurement have increased the complexity of this domain. The recent emergence of the Health Level Seven Fast Healthcare Interoperability Resources (FHIR) standard for patient data representation is particularly pertinent.

Accordingly, the panel members, who are clinicians and co-chairs of the Health Level Seven International (HL7) CDS Work Group, will address two learning objectives. First, attendees will learn key details of extant and proposed HIT standards that are applicable to CDS which they may need to use in their work. Second, the attendees will recognize how these standards have been applied in actual implementations that are used to provide CDS that can be used to improve the quality of clinical care and outcomes.

General Description
The presentation will consist of two themes that align with the learning objectives of the panel presentation. The first theme will address the development of standards for knowledge representation and for related areas such as patient data representation and ontologies. In light of increasing emphasis by regulators and payers on measurement and reporting of clinical quality, more recent activity has included work to consolidate standards for representing and implementing clinical quality measures with standards for CDS. Within the past year, this suite of options for implementing CDS has been expanding with proposed standards to improve representation of clinical quality measures, CDS knowledge and the data needed to drive their implementation.

These standards and proposed standards may be classified into three categories, and the presenters will discuss details of standards in each. Knowledge representation standards provide a way to represent clinical and scientific knowledge in a structured way that can be shared among CDS systems and applied to data to produce CDS. Pertinent standards include the Arden Syntax, the Healthcare Quality Measure Format (HQMF), the Clinical Quality Language (CQL) and the CDS Knowledge Artifact Specification. A second category of CDS standards involves providing access to knowledge without transferring it among implementation sites. Relevant standards in this category include the Decision Support Services (DSS) and Infobutton standards. Finally, a third category of CDS standards provides formalisms for structured representation of data against which knowledge will be applied in order to deliver CDS. This category includes the Fast Healthcare Interoperability Resources (FHIR) standard and its profiles relevant to CDS, the Virtual Medical Record (vMR), the Quality Improvement and Clinical Knowledge Data Model (QUICK) and the Quality Data Model (QDM). State-of-the-art updates on these standards and proposed standards will be provided during this session. The goal is for attendees to recognize how these standards address the key challenges of implementing CDS and how they fit together to facilitate comprehensive solutions for delivering knowledge-driven interventions, emphasizing late-breaking developments as we refine and improve the building blocks for implementing CDS and related quality measures.
The panelists also will describe related efforts in other standards development organizations (SDOs), including the Object Management Group (OMG) and ASTM International, as well as work being conducted via the Joint Initiative Council (JIC) for global health informatics standardization. The context for these efforts, including public policy and collaborative projects that promote CDS, will be described.

The second theme will elaborate how these standards are incorporated in implemented health information systems to facilitate the delivery of CDS and how barriers to the implementation and use of CDS can be overcome. This will provide practical illustrations of how these standards can be used in health care delivery as well as identification of barriers to their use and ways to overcome these barriers. These illustrations will include descriptions of software and projects that leverage CDS standards to deliver knowledge-based interventions in a scalable manner, such as the OpenCDS initiative (www.opencds.org), the OpenInfobutton initiative (www.openinfobutton.org), and commercial implementations of the Infobutton standard. This second theme will be geared particularly toward the clinician who may be contributing expert knowledge to, supervising or using CDS systems in the context of EHRs.

The presenters are co-chairs of the HL7 CDS Work Group, the premier SDO for HIT. They also have experience as clinicians in deploying CDS solutions in a variety of settings. In addition, one of the presenters is the initiative co-coordinator for an effort sponsored by the US Office of the National Coordinator for Health Information Technology (ONC) and the Center for Medicare and Medicaid Services (CMS) to develop and validate a harmonized set of standards for CDS and electronic clinical quality measurement for potential use in US federal regulations. Together, they will present the details of the latest work in HL7 and other SDOs related to CDS, as well as solicit feedback that will help improve the formulation of these standards and their use in practice.

In particular, individual members of the panel will address key aspects of the overall themes:

- **Dr Jenders** will lay the foundation for the panel discussion by describing the overall context of the need for CDS standards. He also will describe the current status and future plans for the Arden Syntax and the Health Quality Artifact Reasoning and Expression Logic standards. He will further review related standards efforts in other SDOs besides HL7.
- **Dr Del Fiol** will describe the latest versions and deployment of the Infobutton standard.
- **Dr Kawamoto** will discuss efforts to harmonize standards for CDS with those standards for clinical quality measurement. He will also address CDS-related standards activity in the Object Management Group (OMG) SDO and deployments of the Decision Support Service standard through the OpenCDS and CDS Collaborative initiatives.
- **Dr Strasberg** will compare and contrast these various standards and review some important factors to consider when deciding which standards to use for different use cases.

**Explanation**

Emphasis on the use of CDS technology by regulators, payers and health systems coupled with new developments in standards in the past year that affect the implementation and use of this technology make the topic of CDS standards particularly timely. Increasingly, clinicians and health care organizations are being required to institute clinical quality improvement measures that leverage health information systems and to report their performance with regard to these measures. Clinicians, administrators and HIT workers are being called upon to implement, use and respond to this technology as part of efforts to measure and improve clinical quality and health outcomes. Standards for CDS facilitate these activities by supporting online access to knowledge and sharing of knowledge for interventions at the time of clinical decision-making. Key drivers of this activity include the regulatory requirement for meaningful use of electronic health record systems in the USA, as well as pressure to reduce the cost of care while improving the quality of care delivery. Indeed, in some cases (e.g., Infobutton for provider access to clinical evidence and individualized patient education), regulations require use of specific CDS standards.

Consequently, these standards are increasingly important. Knowledge about these standards is useful for clinicians and health care organizations that must make decisions regarding what software to buy and how to configure this software as part of their quality improvement activities. Such knowledge is useful for EHR system vendors, who must decide what standards to incorporate into their products. It is equally important for knowledge vendors and...
professional organizations that publish clinical guidelines and that consequently need to know how to represent the knowledge that they make available for use in information systems.

The expected audience for this didactic panel discussion includes
- Clinicians who will use or configure CDS technology;
- Clinical leaders and administrators who will purchase and deploy CDS technology;
- Health information technology workers who will implement and monitor this technology in the context of an overall electronic health record system; and
- Creators and maintainers of knowledge bases, clinical practice guidelines, clinical quality measures and the like.

The panel members lead efforts in the premier international standards development organization for health information technology to create and maintain these standards. They will provide insights into the development of the standards, how they can be used to achieve quality and other goals, and what the plans for additional development of CDS standards will be.

Discussion Questions

- Which standards do I use and how do I use them to improve the quality of clinical care and practice in my own health care organization?
- When shopping for clinical software, what standards should I insist be incorporated in order to comply with regulations and facilitate interoperability?
- Where can I obtain content for my decision support knowledge base?
- What standards and tools facilitate knowledge maintenance?
- What needs for CDS are unmet or not ideally supported by current standards, and how can this situation be improved?

Organizer Statement

All participants have agreed in writing to participate in the panel.

Acknowledgements

The panel organizer, Dr. Jenders, was supported in part by grants UL1TR000124 (NCATS) and 2U54MD007598 (NIMHD) from the US National Institutes of Health.
The Best of Imaging Informatics Research 2015

Panelists: Charles E. Kahn, Jr., MD, MS* and Bradley J. Erickson, MD, PhD*

*Professor and Vice Chairman, Department of Radiology, University of Pennsylvania
*Professor and Vice Chair for Research, Department of Radiology, Mayo Clinic

Abstract

Over the past decade, there have been rapid advances in imaging informatics research. As the storage and display of multidimensional image data has become commonplace and “off the shelf”, imaging informatics has focused increasingly on information not found in pixels and voxels. For example, new ontologies are now available to capture information in the imaging report. New structured data capture technologies and natural language processing capabilities provide discrete data that can be analyzed in concert with other data from the electronic medical record. Likewise, novel quantitative image analysis techniques give us new methods to diagnose clinical conditions and discern their time course. Machine learning methods applied both to images and imaging reports have accelerated our ability to extract meaningful clinical information from images. Several more practical advances, such as standardized workflow and analytic tools have modernized the practice of radiology and other clinical imaging specialties. Advances in imaging informatics research will be reviewed in this session.

Overview

Biomedical informatics spans a spectrum from basic to applied research, and from molecular to population scale. Imaging informatics concerns information at tissue and organ scales. In the past, imaging informatics at AMIA has focused primarily on image processing and analysis. As PACS have become commodities, and as radiologists have recognized the value of structured information for performance improvement and research, the discipline of imaging informatics and the interests of most AMIA attendees have converged. Yet many AMIA members are unable to stay current with the new developments in each sub-area of biomedical informatics. The data processing and visualization tools being developed by imaging informatics researchers will have broad appeal to the AMIA audience. Likewise, the extensive research to extract meaningful data from narrative radiology reports has wide applicability to other aspects of the electronic medical record that contain narrative elements. Thus, we expect this session to appeal to the typical AMIA attendee.

Methods

PubMed search terms for imaging informatics research have a low precision. However, we will perform a PubMed search using search terms “clinical AND imaging AND informatics”, filtered for the previous twelve months and for human research. In addition, we will review the table of contents for the past year of journals that are likely to contain imaging informatics articles, including the Journal of the American Medical Informatics Association (JAMIA), the Journal of Biomedical Informatics (JBI), the Journal of Digital Imaging (JDI), Radiology, and RadioGraphics. These searches will be supplemented by a call for nominations from luminary imaging informatics groups, including the Imaging Informatics Working Group of AMIA, the Radiology Informatics Committee (RIC) of the Radiological Society of North America (RSNA), the Informatics Commission of the American College of Radiology, and the Board of
Directors of the Society for Imaging Informatics in Medicine (SIIM). The bibliographies of the resulting publications also be searched using a “snowball” method.

The presentation will be divided into two sections. Dr. Kahn will present the best of research in ontologies, structured data capture, decision support, and other science focusing on data other than the images themselves. Dr. Erickson will present the best of research focused on imaging processing and image analysis.

**Panel Organizer Statement**

The panelists listed above have agreed to present the material and lead a discussion. A similar session will be presented at the annual meeting of the Radiological Society of North America (RSNA) annual meeting in December, 2015. This session is supported by the RSNA, the Society of Imaging Informatics in Medicine (SIIM), and the AMIA Imaging Informatics Working Group.
User-Centered Methods to Optimize Clinical Decision Support: Examples from Pediatrics with Applicability to All Care Settings

Dean J. Karavite, MSI\textsuperscript{1}, Eric D. Shelov, MD\textsuperscript{12}, Levon H. Utidjian, MD\textsuperscript{12}, Jeremy J. Michel, MD, MHS\textsuperscript{1}, Eli M. Lourie, MD, MBI\textsuperscript{1}
\textsuperscript{1}The Children’s Hospital of Philadelphia, Philadelphia, PA, USA
\textsuperscript{2}Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

Abstract

Unintended consequences of health information technology (HIT) and clinical decision support (CDS) are frequently a result of poor design. Usability problems in electronic health records (EHR) have gained national attention. The need to increase effectiveness and reduce harm is vital and urgent. User-centered methodologies provide an array of methods to design systems around the requirements of the users and are critical to developing HIT that is safe and effective. In a series of vignettes, the panelists will describe applications of user-centered methods during the design of CDS projects in a multispecialty pediatric healthcare system. The panelists will review common user-centered design methods, suggest how to choose methodologies, and discuss both the benefits and challenges of following a user-centered approach in busy clinical settings. Panelists will describe examples of unintended consequences of CDS interventions and user-centered methods to prevent them. The panel moderator is formally-trained in user-centered methods and all panelists are pediatric informaticists with experience in the design and implementation of HIT/CDS interventions.

Learning objectives:

1. Describe system functionality and desired outcomes independent of implementation approach
2. Use low-fidelity mockups and heuristic review in the design of CDS systems
3. Select among potential user-centered methods to support project development

Relevance of User-Centered Methods

There is a prevailing belief that EHRs and CDS systems will improve patient health outcomes, yet there are increasing examples of unintended consequences, sometimes with disastrous results. Although EHR vendors may be able to apply some methods in user-centered design (e.g. cognitive walk through with simulated patients), other methods may be difficult to apply (e.g. directly observing clinical workflow). CDS implementation teams located in health systems are uniquely capable of applying user-centered methods in the design and evaluation of their projects. Individuals who attend this panel will be better equipped with a knowledge of user-centered methods to help increase the success of their projects and ultimately improve health outcomes.

Panel Description

The panelists will describe their experience applying user-centered methods in CDS projects using vignettes from a multi-specialty pediatric healthcare setting. The moderator—a usability expert—will facilitate discussion regarding practical issues that frequently arise in the selection and application user-centered methods in busy clinical settings.

Moderator: Welcome and session overview
Dean Karavite, MS (5 minutes)

Panelist #1: Decision support for neonatal fever
Eric Shelov, MD (10 minutes, followed by 5 minutes discussion)

Dr. Shelov advises on and builds CDS interventions with largely vendor-supplied tools for The Children's Hospital of Philadelphia’s Quality Improvement teams. Tool development can be challenging, with enthusiastic teams often proposing solutions and interventions before completing a proper workflow analysis. These initial ideas, while well-intended, can not only be suboptimal from the perspective of achieving outcomes, but may also introduce risk via unintended consequences. He will describe the five rights-based process he uses when partnering with project leaders to identify the right CDS approach for their improvement goal, minimize alert fatigue, and optimize
proactive decision support. He will review an example of an intervention developed with this approach to decrease length of stay for hospitalized febrile neonates.

Panelist #2: Asthma care plans across the continuum of care
Levon Utidjian, MD (10 minutes, followed by 5 minutes discussion)
Dr. Utidjian will discuss the implementation of a set of asthma management tools across the continuum of care including after hours telephone care, primary care, specialty care, emergency care and inpatient care. Motivated by patient safety concerns, he will describe how he developed use-cases based on clinician workflows and review of common data entry issues to guide implementation of a new set of asthma management tools. He will also discuss unintended consequence of how changing one workflow for data entry to improve one set of QI metrics inadvertently affected another set of performance measures.

Panelist #3: Delivering obesity management expertise in primary care
Jeremy Michel, MD, MHS (10 minutes, followed by 5 minutes discussion)
Dr. Michel will review how he gathered user requirements, used low-fidelity mock-ups and performed a limited usability test while simultaneously programming a functional decision support tool to implement obesity management best practices in busy primary care settings. He will describe unexpected consequences of his decision support intervention, how these issues were handled at the implementation sites, and plans for preemptively addressing these issues in the future. Examples of issues that will be discussed include (1) decreased system performance due to large patient data files, (2) adjustments necessary for site-specific workflows that were not in place at the initiation of the project, and (3) changes to laboratory tests that affected project functionality.

Panelist #4: Aches and pains of influenza immunization reminders in specialty care
Eli Lourie, MD, MS (10 minutes, followed by 5 minutes discussion)
Dr. Lourie will describe his experience designing decision support for antimicrobial stewardship in a pediatric primary care network. He will discuss the difficulties that arise when adding new processes to existing workflows, even when the clinical community is in agreement that the interventions are necessary. This lecture will explore barriers to the project's success, including static vs. interruptive alerts, provider performance feedback, and unintended consequences of the system.

Moderator: Facilitated discussion and closing
Dean Karavite, MS (25 minutes)
The facilitated discussion will explore strategies for selecting the most appropriate methods for typical projects. Obstacles that arise in the application of user-centered methods will be discussed along with recommendations for best practices to overcome these barriers.

Discussion Questions
The following guiding questions will be used to facilitate discussion:
1. Given what you learned from this project, what new innovations or changes do you envision for subsequent years so that project outcomes can approach their target threshold?
2. Which usability methods do you find most useful during the design phase?
3. When there are discrepancies between self-reported user behaviors with actual system usage (i.e. from audit trails), how do you bring that information back to the clinicians?
4. What challenges do you encounter when implementing CDS for Quality Improvement projects and how can you envision addressing them?

Moderator Attestation
The four panelists, Dr. Shelov, Dr. Utidjian, Dr. Michel and Dr. Lourie, have all agreed to participate in this didactic panel.
The Clinical Quality Framework Initiative to Harmonize Decision Support and Quality Measurement Standards: Defined Standards, Pilot Results, and Moving Beyond Quality Improvement

Organizer: Kensaku Kawamoto, MD, PhD, MHS

Panelists
Kensaku Kawamoto, MD, PhD, MHS
Marc J. Hadley, PhD
Tom Oniki, PhD
Julia Skapik, MD, MPH

Panelist Affiliations
a Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
b MITRE Corporation, Bedford, MA
c Intermountain Healthcare, Murray, UT
d Office of the National Coordinator for Health IT, Washington, D.C.

Abstract
The Clinical Quality Framework (CQF) is a public-private partnership sponsored by the Office of the National Coordinator for Health IT (ONC) and the Centers for Medicare & Medicaid Services (CMS) to create a harmonized set of standards for clinical decision support (CDS) and electronic clinical quality measurement (eCQM). At AMIA 2014, experts presented preliminary results from the first year of this multi-year initiative. In this panel, CQF leaders will describe the standards developed, including the HL7 Clinical Quality Language standard for logical expression, the HL7 FHIR Quality Profiles for data exchange, the HL7 Quality Improvement and Clinical Knowledge model for logical inferencing, and several implementation guides leveraging these foundational standards. The panelists will then present findings from pilot deployments of the standards in various clinical domains ranging from cardiology to radiology, preventive care, infectious diseases, and oncology. The panelists will also describe partnerships with relevant initiatives beyond the quality domain, including the ONC Data Access Framework initiative, the Health Services Platform Consortium (HSPC), and the Clinical Information Modeling Initiative (CIMI).

Description
The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 min</td>
<td>Skapik</td>
<td>Federal government perspective on motivation and goals for CQF</td>
</tr>
<tr>
<td>15 min</td>
<td>Hadley</td>
<td>Standards description</td>
</tr>
<tr>
<td>20 min</td>
<td>Kawamoto</td>
<td>Pilots and partnerships</td>
</tr>
<tr>
<td>15 min</td>
<td>Oniki</td>
<td>HSPC and CIMI collaboration</td>
</tr>
<tr>
<td>30 min</td>
<td>Kawamoto</td>
<td>Panel discussion with audience</td>
</tr>
</tbody>
</table>

Dr. Kawamoto will serve as the moderator and introduce each of the panel members and their organizations. Dr. Skapik will describe the motivation and goals underlying the federal
government’s sponsorship of the CQF initiative. Dr. Hadley will summarize the standards developed by the CQF initiative. Dr. Kawamoto will describe pilot implementations of the standards, as well as partnerships with interoperability initiatives beyond the quality domain. Dr. Oniki will discuss the implications of the initiative beyond the quality domain. These presentations will be followed by a panel discussion with the audience moderated by Dr. Kawamoto.

**Skapik:** Dr. Skapik is the ONC lead of the CQF initiative. She completed her MD and MPH at the Johns Hopkins Schools of Medicine and Public Health, and her residency in Internal Medicine at the University of Pittsburgh Medical Center, joining the faculty in 2011. She is a former AAAS Science and Technology Policy fellow, working at the National Science Foundation in the Directorate for Computer & Information Science & Engineering working on their Smart Health and Wellbeing initiative. Dr. Skapik will discuss the motivation behind the federal government initiating and sponsoring the CQF initiative.

**Hadley:** Dr. Hadley is Senior Principal Software Systems Engineer at MITRE Corporation, an expert on eCQM, and Technical Lead for the CQF initiative. Dr. Hadley will describe the standards that have been developed through the CQF initiative, which include the HL7 Clinical Quality Language standard for logical expression, the HL7 FHIR Quality Profiles for data exchange, the HL7 Quality Improvement and Clinical Knowledge model for logical inferencing, and several implementation guides leveraging these foundational standards. The standards developed and validated through CQF will be available to policy makers for potential inclusion in future federal requirements relating to health information technology.

**Kawamoto:** Dr. Kawamoto is Associate Chief Medical Information Officer, Assistant Professor of Biomedical Informatics, and Director of Knowledge Management and Mobilization at the University of Utah. Dr. Kawamoto is co-Initiative Coordinator of the CQF initiative. Previously, he served as the Initiative Coordinator of the ONC Health eDecisions initiative, which developed the CDS standards being harmonized with eCQM standards in the CQF initiative. Dr. Kawamoto will present findings from pilot deployments of the standards in various clinical domains including cardiology, radiology, preventive care, infectious diseases, and oncology. Dr. Kawamoto will also describe partnerships with relevant initiatives beyond the quality domain to broaden the stakeholder community and reduce implementer burden. In particular, Dr. Kawamoto will describe a partnership with the ONC Data Access Framework (DAF) initiative, which is developing HL7 data interchange standards for Meaningful Use. He will also provide an overview of partnerships with the Health Services Platform Consortium (HSPC) and the Clinical Information Modeling Initiative (CIMI), which are ongoing semantic interoperability collaboratives with a scope beyond quality improvement.

**Oniki:** Dr. Oniki is Senior Medical Informaticist at Intermountain Healthcare and is the coordinator of HSPC and CIMI’s collaborations with the CQF initiative. Dr. Oniki will provide an overview of HSPC and CIMI and its goal to achieve more efficient and effective health care enabled by true semantic interoperability. Dr. Oniki will discuss the need for coordinated, isosemantic data models across interoperability initiatives, so as to avoid the proliferation of incompatible standards, reduce implementer burden, and maximize the adoption of a common set
of standards. He will then provide details on how HSPC and CIMI are collaborating with the CQF initiative.

Kawamoto: Dr. Kawamoto will lead a moderated discussion with the audience. The objectives of this discussion will be to answer questions from the audience and to obtain feedback from the audience to guide the future direction of standards development and implementation in the quality domain. The discussions questions listed below will be used to stimulate this discussion.

**Significance of panel topic and anticipated audience**

Current and proposed federal regulations include requirements for divergent CDS and eCQM standards. The CQF initiative has addressed this important challenge by developing and validating a unified set of standards for CDS and eCQM. Of note, the results of the CQF initiative will be available to federal policy makers for potential inclusion in future federal regulations. Therefore, the panel topic is timely, urgent, needed, and of great significance to the clinical informatics community. The anticipated audience is AMIA attendees interested in CDS, eCQM, data model standards, other interoperability standards, and/or federal regulations in this area, including members from academia, the public sector, and private industry.

**Discussion questions**

What do you think about the proliferation of standards and efforts to coordinate among standards development initiatives? Do you think the alignment efforts described are headed in the right direction?

What recommendations do you have for standards development and implementation in this area?

What gaps still exist in the available standards related to CDS and eCQM? How should we address those gaps?

What clinical domain areas can most benefit from standards-based, interoperable CDS and eCQM (e.g., immunizations, chronic disease management, chemotherapy)?

How should CDS and eCQM be leveraged in the short-term and long-term to improve clinical quality and health outcomes?

**Participation statement**

All proposed panelists have agreed to participate in the panel if the proposal is accepted.

**Acknowledgements**

The CQF initiative is supported by funding from the ONC and CMS, as well as by voluntary contributions of effort from various members of the CQF community.
Informatics Approaches to Supporting Emerging Accountable Health Care Delivery Models

Gilad J. Kuperman, MD, PhD (1), David W. Bates, MD, MSc (2), David C. Kaelber, MD, PhD, MPH (3), David A. Dorr, MD MS (4)
(1) NewYork-Presbyterian Hospital and Columbia University, New York, NY, (2) Brigham and Women’s Hospital, Partners HealthCare System, Harvard University, Boston, MA, (3) The MetroHealth System, Case Western Reserve University, Cleveland, OH, (4) Oregon Health and Science University, Portland, OR

Abstract

Medicare, Medicaid and commercial payers are creating incentives for efficient care delivery models and the health care market place is responding rapidly. The new incentives require health care delivery organizations – often for the first time – to include efficiency of care delivery as a key strategic goal. These dynamics are changing the way provider organizations are delivering care. New information infrastructures -- including analytics, work flow support for new care roles (e.g., care coordinators), information systems support for team-based work flows, information-enabled interactions with new business partners, and improved interactions directly with patients -- are emerging to manage patients in the new environment. The learning objective for this panel is for attendees to develop a better understanding of how leading organizations are addressing changes in health care payment models and the informatics-enabled approaches they are taking to support workflow changes, team-based care and the need for data-driven organizational and health care management.

General description.

The panelists are leaders at large organizations that are developing informatics-based approaches to the new health care environment. Each panelist will describe the following topics from the perspective of his organization: (i) the business drivers that are requiring health care organization to consider new, efficiency-based, approaches to care, (ii) the new models of care delivery that are being established, (iii) the information technology requirements emerging to support the new models of care, (iv) the current state of implementation, (v) challenges being faced and approaches to overcoming those challenges, and (vi) the role of informatics in the new environment. Dr. Kuperman will serve as the panel moderator and also as a panelist. This will be an “option 1 (didactic)” panel – each panelist will speak for 10’ with 5’ for questions and there will be 30’ at the end for audience-based discussion.

Description of each panelist’s presentation

Gilad Kuperman, MD, PhD, is Director for Interoperability Informatics at NY- Presbyterian Hospital (NYPH), a 2400-bed, $4 billion, organization in New York City. NYPH is participating in New York State’s $6 billion Delivery System Reform Incentive Program (DSRIP).

The goal of DSRIP is to reduce avoidable admissions and emergency department visits by 25% over a 5-year period. DSRIP is creating incentives for provider organizations to more effectively engage non-acute, social service and community-based organizations in the care of patients that consume large amounts of health care resources. The engagement of these organizations will allow the patient to remain in the community more effectively and avoid the use of costly emergency and inpatient-based services. Examples of such organizations include mental health, substance abuse, housing, nutritional, employment, home health, skilled nursing and other services. NYPH has engaged over 70 such organizations as part of its DSRIP project.

Care models have been designed that will rely on the ability to identify the relevant high risk patients, connect the patient with a care coordinator, and engage the relevant support service organization. The information systems requirements analysis has identified the need for (i) analytics support for cohort identification, management and quality reporting, and predictive modeling, (ii) workflow support for care managers, for example clinical documentation, tasking and roster management, (iii) health information exchange capabilities to support sharing of
clinical data among participants, (iv) enhancements to the electronic health record, such as alerts and worklists, (v) a patient portal that supports communication between patients and providers, and (vi) and application to support the work flow of community health workers. To support the program, NYPH will be leveraging and extending its existing enterprise analytics platform, its care coordination workflow application and its electronic health record. NYPH will be working with its regional health information organization (RHIO) to support health information exchange among the partners. The project has been in the planning stages since the Fall of 2014 and officially launches in April, 2015.

David Bates, MD, MSc, is the Chief Innovation Officer at Brigham and Women’s Hospital (BWH), which is a 700-bed, $1 billion/year organization which is part of the Partners Healthcare System, the largest integrated delivery system in New England. Currently, of patients at the BWH and in Partners, about 30% are in some type of ACO arrangement—nationally, about half the ACOs are in Massachusetts.

As a result, BWH and Partners are implementing an array of approaches to being to more effectively manage populations. For our current ACO populations, we risk stratify all of them, and have aggressively managed the highest-risk group, which has resulted in substantial decreases in the trend toward increasing total medical expenditures for this group. We are developing an Enterprise Data Warehouse (EDW), which will include data from both our Epic clinical systems, but also from a number of other sources, including an all-payer claims database. In addition, we are extending our analytics tools, attempting to develop our electronic tools around care coordination, developing our registry capabilities, and refining our ability to perform cost accounting. All of these areas will be discussed.

David Kaelber, MD, PhD, MPH is the Chief Medical Informatics Officer and Vice-President of Health Informatics at the MetroHealth System, a $1 billion/year integrated healthcare delivery network which includes a 500-bed hospital and over 20 outpatient clinics throughout northeast Ohio. The MetroHealth System is the public healthcare system for the greater Cleveland area and is a Medicare Shared Saving ACO. The MetroHealth System was also the first public healthcare system to begin to install Epic in 1999 in the outpatient setting and to achieve HIMSS Electronic Medical Record Adoption Model Stage 7 hospital and ambulatory recognition with the Epic electronic health record in 2014.

David Kaelber is responsible for all health informatics, patient engagement, and population health activities in the MetroHealth System. In support of the System's ACO efforts the MetroHealth System has developed a number of reporting registries, risk stratification tools, and interventions to improve our care and engage our patients more, especially outside of face-to-face visit. Specific informatics effort have included developing and implementing registries integrated into clinical work flows, deploying ACO metric dashboards, constructing risk stratification tools within our EHR, integrating payer data into our clinical data repository, maximizing use and functionality of our personal health record, and implementing technology to automatically call and text patients. All of these areas (and more) will be discussed.

Some specific examples to be discussed include:

1. ACO dashboards – how we have used “out of the box” vendor MSS ACO provider level and system level dashboards to track and improve ACO performance.
2. Informatics outreach interventions to improve ACO metrics - how we have used registries, bulk ordering, bulk messaging, and self-scheduling to double mammography screening rates in patients behind on mammograms.
3. Claims data – how we integrated claims influenza data to improve clinical ACO immunization metrics.

Dr. Kaelber is also on the Board of CareSource, the largest managed care Medicaid insurer in OH, and so will discuss informatics approaching to accountable care from the payer perspective as well.

David A. Dorr, MD MS is Associate Professor and Vice-Chair at Oregon Health and Science University. He studies how information systems, incentives, and practice facilitation can improve health care for our most complex patients while reducing costs.

In Oregon, the business drivers for efficient care models include an increasing prevalence of population-based payments, for example, from coordinated care organizations (CCOs) and through initiatives such as comprehensive
primary care from the Center for Medicare and Medicaid Services (CMS). To address the goals of these programs, the provider organization is giving increasing emphasis to care coordination, population management services, and risk stratified care management. The use of information technology has varied across initiatives; results from two trials that combined HIT with technical assistance will be described. Technology, while implemented, requires constant revision to meet program needs, and informatics is at the center of the difficult aspects.

**Why this topic is timely**

Accountable models of care that encourage efficiency of care are increasing in prevalence. In January, 2015, Secretary of HHS Silvia Burwell announced that HHS has a goal of tying 30% of HHS payments to alternative payment models “such as accountable care organizations” by the end of 2016 and 50% by the end of 2018. Medicaid and commercial payers similarly are changing their payment models. These changes require health care organizations to be more effective and efficient. These dynamics require health care providers to use their information systems in new ways that go far beyond the routine transactional capabilities of certified EHRs. Leading organizations are learning about the best way to support team-based care, new provider roles, risk-based approaches, patient engagement and health information exchange. The topics that will be covered in this panel will be very appealing to the AMIA Annual Symposium attendees that is interested in clinical informatics and the application of new and established technologies in the new environment.

**List of discussion questions to enhance audience participation**

1) What are the limitations of the current generation of EHRs in supporting the new emerging care delivery models and what additional, supportive or “wraparound” technologies are needed to meet the emerging requirements?
2) To what extent do the information technology requirements change over time as the organization matures in its experiences with accountable care models?
3) To what extent are patient facing technologies a critical part of the infrastructure to support accountable care?
4) Are mHealth or other innovative information technology approaches helping to support these new care delivery models?
5) What are the strengths and weaknesses of the current generation of readmission and resource utilization risk prediction models?
6) To what extent are the workflows for team-based care well-understood and easy to implement? How does integration among people, process and technology evolve over time?
7) How effective are current health information exchange strategies at supporting these new models of care and where is improvement needed?

**Statement from panel organizer that participants have agreed to take part.**

I, Gil Kuperman, attest that the other panel participants named here have agreed to participate.

**Bibliography**

Collaboration and Health Information Technologies: Towards Defining and Operationalizing the Collaboration Space

Craig Kuziemky, PhD 1, Madhu Reddy, PhD 2, Katie A. Siek, PhD 3, Sarah Collins, PhD, RN 4
1University of Ottawa, Ottawa, ON, Canada; 2 Northwestern University, Evanston, IL, USA; 3 Indiana University, Bloomington, IN, USA, 4 Partners Healthcare Systems, Boston, MA

Abstract

Despite calls for increased collaborative care delivery it is still a challenge to operationalize collaboration, and more specifically, to design and evaluate HIT to support collaboration. Reasons for that include a lack of integrated studies on different aspects of collaboration and a lack of research with an explicit focus on collaboration. However, studying collaboration can be challenging given the range of processes, providers and settings where collaboration takes place. To advance research on collaboration and HIT we need ways to bound studies of collaboration according to the different type of collaborations and the overall collaboration space where they exist. This panel will provide insight on how to define and operationalize the collaboration space. We will discuss the structure of the collaboration space and how it provides bounding for studies of collaboration as well as discuss clinical collaboration and patient centered participatory medicine as two specific contexts of collaboration spaces.

Introduction

Collaboration can be defined as ‘planned or spontaneous engagements that take place between individuals or teams of individuals, whether in-person or mediated by technology, where information is exchanged in some way (either explicitly, i.e. verbally or written, or implicitly, i.e. through shared understanding of gestures, emotions, etc.), and often occur across different roles (i.e. physician and nurse) to deliver patient care’ [1]. There have been recent calls for transforming the healthcare system to one that is collaborative across settings, diseases and governance structures [2]. While reports and commissions in many different countries have called for increased collaboration in the healthcare domain, operationalizing collaboration is a significant challenge because of the need to integrate various providers, processes, technologies and settings [3, 4]. We have seen a growing body of research on various aspects of collaboration including education [5], teamwork [6,7], patient-centeredness [8], technologies’ impact on collaboration [9], clinical content development [10], and designing and evaluating for collaboration [11]. However, a significant shortcoming is that these efforts have looked at isolated aspects of collaboration rather than the overall space where collaboration exists. Furthermore, collaboration is often only implicitly discussed in these studies rather than being the focal point of research.

Consequently, the lack of explicit focus on collaboration has led to challenges because of the lack of fit between HIT and the collaborative contexts where it is used [12]. Studies have shown that simply providing tools to support collaboration without facilitating the necessary behavioral support will result in less than ideal collaborative outcomes [13]. While there is a need to explicitly study collaboration, a challenge is that the collaboration space and the concepts associated with it are so varied that it is difficult to identify how to ‘bound’ studies of collaboration. The collaboration space refers to the technologies, contexts, processes and outcomes that are involved in collaboration [14]. To advance research on collaboration and HIT, we need to not only operationalize the concept of collaboration but also examine the different types of on-going collaborations that shape the collaboration space, and the role of HIT in supporting these various collaborations. We also need to look at how the structure of the collaboration space would be operationalized by different behavioral contexts of collaborative care delivery.

Given the calls for increased collaborative care delivery we believe it is important to engage in active dialogue about how collaboration currently occurs and how we can make it better. This panel will start to address that need identified needs by explicitly examining collaboration spaces, the role these spaces play in healthcare delivery, and how biomedical informatics and HIT can better support different collaboration spaces. This panel will bring together panelists with a wide variety of perspectives on collaboration and HIT from design, implementation, and use perspectives. Our goal for the panel is for each panelist to very briefly discuss their particular topic and then use the rest of the session to engage the audience in the issues of investigating and designing for collaboration.
Learning Objectives:
Attendees will be able to:
- Understand the current state of research on collaboration and the need to explicitly study collaboration
- Appreciate the challenges in operationalizing the collaboration space
- Identify ways of bounding the collaboration space to enable more focused studies of it
- Realize how different behavioral contexts influence the collaboration space

Discussion questions that will be posed to the audience include:
1. What are some components of collaboration that we need to focus on in research studies?
2. What are the most useful methods to capture collaboration?
3. How do different contexts of healthcare delivery impact collaboration?
4. To what extent does different HIT application support collaboration? How can HIT be better designed to support collaboration?

The target audience is HIT designers, clinicians and researchers with an interest in collaborative healthcare delivery.

Panelists and Their Insights on Operationalizing the Collaboration Space

The four panelists will each provide unique perspectives on how to operationalize collaboration.

Craig Kuziemsky, PhD – Dr. Kuziemsky has been studying the modeling, design and evaluation of HIT to support different collaborative contexts. His work has focused on the structural aspects of collaboration as well as behavioral aspects that shape the structure such as common ground. Dr. Kuziemsky is also interested in the different contexts in which collaboration occurs and how these contexts influence HIT design and evaluation. In his presentation Dr. Kuziemsky will focus on behavioral aspects of the collaboration space and the need to consider the relationship between structure and behavior in designing for collaboration. He will also discuss the need to consider trade-offs at the individual-collaborative interchange.

Madhu Reddy, PhD – Dr. Reddy has been studying collaboration and HIT from a sociotechnical perspective for the past fifteen years. He is particularly interested in the design, implementation, and adoption of healthcare technologies in highly collaborative and information-intensive clinical settings. In his presentation, Dr. Reddy will focus on ways on how we can “bound” collaboration so that researchers can focus on the actually collaborative aspects in a study. He will also discuss ways to operationalize the collaboration space.

Katie Siek, PhD – Dr. Siek has studied the interface between everyday people and clinicians through innovative wearable and mobile systems that connect with personal health record systems for over ten years. She is particularly interested in the validity of data transferred between groups and the implications of that data transfer on the overall patient-provider relationships. In her presentation, Dr. Siek will discuss some of the technical and social issues within the broader collaboration space. She will also highlight the implications of lay individuals sharing personal health management data with clinical collaborators.

Sarah Collins, RN PhD – Dr. Collins will focus on the collaboration spaces within inpatient clinical care units with a particular focus on tools to support interprofessional collaboration in critical care and patient safety. Critical care collaboration relies on a blend of formal and informal communication channels requiring reliable tools to exchange information (including safety concerns) and flexible tools to establish common ground for complex and patient-tailored or situation-specific clinical concepts. This presentation will describe the patient safety and patient-centered care planning dependencies woven through the interactive communication spaces of critical care interprofessional rounds [15], profession-specific handoff, profession-specific planning and decision making [16], and transitions of care. Specific tools that have been designed, developed, configured, integrated, and evaluated to support these dependencies at Brigham and Women’s hospital will be discussed, including challenges and remaining barriers. These tools include electronic plans of care [17], rounding tools, interprofessional safety checklists, communication microblog, patient facing safety reporting, and clinical dashboards.

Dr. Kuziemsky will also serve as moderator for the panel. All participants have agreed to take part on the panel.

References
The Implementation of Online Patient Portals in Safety Net Settings: The Realities of Meaningful Use Certification with Vulnerable Patient Populations

ABSTRACT
This panel will provide insights and early data about the experiences of safety net healthcare settings implementing online patient portals in order to meet Meaningful Use Stage 2 certification deadline during 2015. While the vast majority of US healthcare systems are in the midst of implementing patient access to online patient portal websites to allow patients to view their medical information online and communicate electronically with providers in between office visits, healthcare systems serving predominantly low-income and uninsured/publicly insured face unique challenges with patient as well as provider/staff engagement. In particular, these systems have developed strategies for engaging patients with limited health literacy and/or basic computer/Internet skills, such as dedicated portal training programs. They also have recognized the need for parallel processes for provider/staff engagement.

Learning objectives:
- Determine key lessons learned for developing support services/programs to support vulnerable patients in portal use, and present real-world Stage 2 Meaningful Use metric data from several safety net systems
- Identify policy and practice changes (specifically for Meaningful Use criteria) that might alleviate provider resistance to adopting secure messaging, especially in resource-constrained healthcare settings
- Formulate an approach to communicating with EHR/portal vendors when implementing a product that must meet the needs of safety net patients and providers (e.g., literacy, language standards)

PANEL
This panel will include leaders from both an urban, academic-affiliated safety net healthcare system as well as a community health center serving a rural population in California – with perspectives from research, practice, and leadership roles. The panel organizer (Dr. Lyles) is a health services researcher based at UCSF in the Center for Vulnerable Populations, who has been studying portal uptake across several healthcare systems in the past 5 years. Dr. Sarkar is practices internal medicine at General Medicine Clinic at San Francisco General Hospital, an urban public hospital and academic medical center. Dr. Ratanawongsa is the Chief Medical Informatics Officer for the San Francisco Health Network, which is a consortium of San Francisco Department of Public Health hospital- and community-based primary care clinics located across the city. Finally, Dr. Oryn is the Chief Medical Informatics Officer of an independent health center in Petaluma, California serving approximately 20,000 low-income and predominantly uninsured patients.

In 2015, we in the medical informatics community will continue to spend a significant amount of time discussing electronic health record (EHR) and portal implementation. This is particularly true as a large proportion of healthcare systems will moving from basic EHR implementation to patient engagement with portals in order to meet Stage 2 of Meaningful Use criteria to receive federal incentive dollars. While this huge increase in patient engagement EHRs is likely to revolutionize care in many ways, it is unclear how it will influence existing healthcare disparities for more vulnerable patient populations. This panel will shed light on this less-discussed, but critical issue, including advocating for clear next steps for research, practice, and policy for portal uptake among diverse populations – particularly related to literacy and language standards.

1. PATIENT PERSPECTIVES
Courtney R. Lyles, PhD (panel organizer)
Assistant Professor
University of California, San Francisco
Center for Vulnerable Populations

Summary:
Over the past year, Dr. Lyles has conducted in-depth interviews with patients at San Francisco Health Network, including 1) open-ended discussions about technology use and their ability/interest in using a portal website for healthcare management, and 2) thinkaloud interviews to assess usability and accessibility of the existing portal. San Francisco Health Network includes 11 community-based primary care clinics as well as 4 hospital-based primary care clinics at San Francisco General Hospital.
**Specific Findings to Date:**

We have conducted 16 in-depth interviews with chronic disease patients or their caregivers to date (with thinkaloud/usability interviews ongoing). The patient sample included 11 patients and 5 caregivers who were existing computer users. The mean age was 55 years, 50% were African American, 63% were male, and 70% reported using the Internet daily – although the range of computer abilities/tasks was broad, with a third of participants not using email.

The in-depth patient interviews focused on experiences with the healthcare system, technology use in everyday life, and interest in using an online portal to manage healthcare tasks. The major barriers to portal use were: concerns about security of information online, lack of technical skills/interest, and preferences for in-person visits/communication. Several participants discussed fundamental barriers to using an online portal, including: challenges with reading, writing, and typing (“I think that’s the reason why I don’t really use the computer a lot because it’s a lot of reading. Like I said, I’m not really a heavily educated guy.”); personal experiences with online security breaches/viruses; and a distrust of potential portal security measures (“Regardless of what a person says that this site is secured and all that, I just don’t believe it.”). Overall, participants also saw the value of a patient portal to increase the convenience of care coordination (“Instead of bugging the front desk…and they have to look everywhere for you…it’s convenient I think just looking at the calendar yourself.”) and support patient-driven communication (“I just hope that it’s just useful when you’re not at the doctor and you just want to know different information or contact your doctor.”).

2. **PROVIDER PERSPECTIVES**

Urmimala Sarkar, MD, MPH
Division of General Internal Medicine at San Francisco General Hospital
San Francisco, CA

**Summary:**

Dr. Sarkar has been a primary care provider in the General Medicine Clinic at San Francisco General Hospital for the past 8 years. Her presentation will focus on the provider experience in implementing a patient portal. This will include: 1) results from provider and staff survey and focus groups in the San Francisco Health Network that identified the major barriers and facilitators to portal use as a part of standard care, and 2) a qualitative examination of the content of early provider adopters of secure messages in this setting. This presentation is specifically timely given the direct link between patient engagement and provider/staff engagement, as previous evidence has demonstrated that one of the key successes to portal uptake among patients is the expectation that these websites extend existing provider relationships.

**Specific Findings to Date:**

To date, we have completed a survey of 77 primary care providers and staff (January and February 2015), assessing interest in recommending portal use to patients and expected barriers in using the system, particularly in using secure messaging through the portal. The sample included 46 physicians, 18 other providers (nurse, medical assistant, etc.), and 11 staff members, and about half of whom (45%) were from community-based clinics. The mean age was 42 years, and 77% were female and 46% were white. Overall, 45% said they would recommend the portal to most of their patients, and 38% said they would recommend the portal to some of their patients. Viewing upcoming appointments and lab results were the features seen as being recommended the most often. In terms of challenges for their patients in using a portal, English proficiency (81%), literacy (75%), unreliable access to computers/Internet (71%), and unstable housing (61%) were the biggest reported perceived barriers for patients. When considering using secure messaging through the portal their own practice, lack of time (68%) was the biggest barrier, followed by worry about patient overusing or sending irrelevant emails (47%).

3. **LEADERSHIP PERSPECTIVES**

Neda Ratanawongsa, MD, MPH
Chief Medical Information Officer, San Francisco Health Network
San Francisco, CA

**Summary:**

Dr. Ratanawongsa serves as the Chief Medical Information Officer for the San Francisco Health Network, which includes the outpatient clinics based at San Francisco General Hospital (such as General Medicine Clinic) as well as
the 11 community-based public clinics throughout the city and county of San Francisco. Her presentation will focus on the administrative and leadership experience in implementing an online portal across multiple clinic sites. Specifically, she will discuss the considerations for portal implementation given 1) the existing information technology infrastructure in safety net systems and 2) the realities of portal implementation given the linguistic and cultural diversity in an urban public healthcare setting.

Specific Findings to Date:
The San Francisco Health Network has developed several strategies for implementing a portal that will be most accessible to their population. This includes:

- A dedicated Patient Engagement Workgroup that includes membership from inpatient, outpatient, administration/leadership, quality improvement, and information technology teams/departments
- A partnership with the San Francisco General Hospital Library to be able to train patients how to 1) use computers for basic tasks, and 2) specifically enroll and access features of the MySFHealth portal website
- There have been multiple delays and challenges in working with the EHR vendor to make the portal website more accessible. The San Francisco Health Network is still waiting to be able to pilot a Spanish version of the product, as it is not currently available.
- The system has had to develop a robust plan for proxy access to ensure patients with limited English proficiency have the opportunity to dedicate a family member to access the portal website if appropriate

4. REFLECTIONS FROM A COMMUNITY-BASED CLINIC
Danielle Oryn, DO
Chief Medical Information Officer, Petaluma Health Center
Petaluma, CA

Summary:
Dr. Oryn is a practicing clinician and CMIO of the Petaluma Health Center in Northern California. This community-based clinic system includes 44 providers serving 32,000 patients in 3 clinics in Sonoma County, California. Her presentation will cover the range of patient, provider, and leadership experiences in implementation of the portal in this healthcare system serving primarily uninsured or publicly uninsured patients in a more rural area of the state.

Specific Findings to Date:
In this community setting, Petaluma has been an early leader in successfully registering their patient population for portal use, with 44% of active adult patients (seen in the last year) currently web-enabled. The key lessons learned to date include:

- Patients need help getting on to the system for the first time. Each clinic has had to create specific workflows to assist patients with registration.
- Language and literacy issues continue to be a struggle.
- Minors and proxy settings are not adequate with the EHR vendor – each clinic has also had to create policies and workarounds for this.

DISCUSSION QUESTIONS

- Do you think there should be specific Meaningful Use requirements for literacy and language access to portals?
- What are some steps that can be taken to alleviate provider resistance to adopting secure messaging, especially in resource-constrained healthcare settings?
- What are the differences in portal implementation challenges between urban and rural settings?
- How can vulnerable patient populations’ experiences of online portal use be improved?
- Please discuss your experiences delivering portal access (or other mHealth solutions) via mobile phones in your settings.

PARTICIPATION
All participants have agreed to take part on this panel during the AMIA conference in November 2015 in San Francisco, CA.
Health information technology and large-scale adverse events

Farah Magrabi PhD¹, Dean F. Sittig, PhD, FACMI², Jean M. Scott³, Peter M. Kilbridge, MD⁴.

¹Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia; ²University of Texas – Memorial Hermann Center for Healthcare Quality & Safety, School of Biomedical Informatics, University of Texas Health Sciences Center, Houston, Texas; ³Informatics Patient Safety, Office of Informatics and Analytics, Veterans Health Administration, Albany, NY, USA; ⁴The Advisory Board Company, Washington, DC, USA.

There is an urgent need to address large-scale adverse events associated with health information technology (HIT) because the opportunity for harm to numerous patients is intensified with rapid implementation of systems worldwide. This panel will review current evidence about large-scale events including their impact on care delivery and consequences for patients. It will then use a case study of a high profile failure to discuss lessons for safe implementation and operation of present day HIT systems. The panel will then turn its attention to processes for detection and management of HIT events. And finally it will examine strategies for prevention and mitigation of large-scale events. The panel will aim to improve understanding about large-scale events and to transfer learnings about evidence-based best practice methods to improve preparedness, detection, response and recovery from HIT events, and to identify gaps and areas for further work.

Introduction

In February 2015, a Northern California hospital lost access to its electronic health records (EHR) for a week¹. At around the same time, hundreds of patients were exposed to the risk of harm for more than 14 hours when all the computer systems crashed at the Fiona Stanley Hospital, Australia’s most technologically advanced facility². A few months earlier, another health information technology (HIT) related event – this time in the state of Queensland – disrupted care delivery to thousands of patients when vital clinical systems were not available to doctors for several hours in every hospital across the state due to a data centre outage³. Another data centre outage affected 80 public sector organisations called trusts within England’s national health service (NHS)⁴. And in the US Veterans Health Administration, it took more than nine hours to restore services to the 17 sites in that were affected by an outage⁴.

There is now compelling evidence that such HIT events can introduce new, often unforseen, errors which can affect the safety and quality of clinical care and may lead to patient harm⁵. Like events associated with faulty equipment, HIT-related events because of their scale and scope can instantly mushroom into large-scale adverse events where individual events harm or increase the risk of harm to numerous patients⁶. There are few classes of intervention where risks to patients are rapidly increasing in scope. The opportunity for patient harm is intensified with growing numbers of large and complex HIT systems being deployed rapidly at organizational, regional and national levels to reform health systems worldwide⁷. For instance, 94% of US hospitals used EHRs in 2014, a five-fold increase since 2008⁸. Yet there is no active surveillance of the frequency and scope of adverse events involving HIT. A 2014 survey of US hospitals, found that large-scale events lasting 8 or more hours were a common phenomenon⁹.

This didactic panel will focus on large-scale adverse events associated with HIT. The types of technical problems and human factors issues contributing to large-scale events will be reviewed. The panel will then examine a case study of a high profile failure as well as processes at the Veterans Administration for detection and response to HIT events. The final presentation will examine strategies to proactively minimise the risks of large-scale events through processes for system design, implementation and use. The panel will be moderated by Farah Magrabi.

The problem: large-scale adverse events associated with HIT

Organiser and presenter: Farah Magrabi

This presentation will introduce the topic and review current evidence about HIT events as well as their propensity to lead to large-scale adverse events. In addition to the published literature it will draw upon a database of events...
including voluntary reports to the US Food and Drug Administration (FDA) and all the events over a 6-year period from England’s national program for IT to examine the nature and consequences of large-scale events\textsuperscript{16,13}. It will then give an overview of the types of technical problems and human factors issues as well socio-technical contextual variables contributing to events (e.g. training, cognitive load and clinical workflow). And finally, it will discuss the challenges in measuring the effects of large-scale events on care delivery and patient outcomes.

**Case study: Lessons from a large-scale failure**

*Presenter: Peter M. Kilbridge*

*This presentation will examine a large-scale failure* at the Beth Israel Deaconess Medical Center in Boston where it required four days to restore systems following a network disruption\textsuperscript{12}. It will specifically discuss the lessons that can be learnt from this event for safe implementation and operation of present day HIT.

**Detection and response to HIT events at the Veterans Health Administration**

*Presenter: Jeanie M. Scott*

*This presentation will provide an overview of the management of HIT events at the Veterans Health Administration.* Building upon standard processes for IT service management HIT events posing actual and potential risks to patients are explicitly managed by a dedicated patient safety team\textsuperscript{15}. The Informatics Patient Safety office is the focal point for the systematic and strategic evaluation of HIT safety risks through collaboration with the National Center for Patient Safety and IT organizations. This presentation will describe several examples including changes in HIT deployments\textsuperscript{16,17}. The main focus will be to identify practices during deployments to minimize risks.

**Improving HIT safety through design, implementation and safe use**

*Presenter: Dean F. Sittig*

In an attempt to create a set of proactive, self-administered EHR risk assessment tools, the US Office of the National Coordinator for Health Information Technology sponsored development of the Safety Assurance Factors for EHR Resilience (SAFER) guides\textsuperscript{9}. The SAFER guides are based on an extensive, human factors-type informatics research project conducted over a 2-year period. This research project identified six key principles that guided development of the SAFER recommendations. Data and systems are:

1. Available when and where needed
2. Only viewed by authorized users
3. Only modified by authorized users
4. Used correctly and completely throughout the organization
5. Designed and implemented to promote safe, effective, and efficient use
6. Used to monitor, detect, and report on the safety of the system

*This presentation will describe how the SAFER guides can be used to help healthcare organizations assess the safety and effectiveness of the EHRs as they are used in the clinical setting.* The main focus will be on best practices for preventing and mitigating large-scale events involving EHR downtime.

**Conclusion**

The presentations will show that events involving large and complex HIT systems are routinely posing risks to increasing numbers of patients. Participants will gain an insight into the types of technical and human factors issues that give rise to large-scale adverse events and their impact on patient safety. The panel will showcase practical strategies for improving the detection and management of HIT events and for addressing these issues throughout the HIT lifecycle. Participants will see that most safety problems with HIT are amenable to mitigation and amelioration with careful system design, implementation and use.

**Strategy to engage the audience**

At the conclusion of the presentations the moderator will invite the audience to participate by asking questions to various panel members or by voicing their views and opinions. If required the audience will be presented the following list of questions to stimulate discussion:

1. What do you think is the biggest safety risk for EHRs?
2. What is the most important thing an organization could do to reduce their safety risks?
3. What are the main challenges for researchers and practitioners to improve the safety of HIT?
4. How can we improve monitoring of the frequency and scope of large-scale adverse events?
5. What strategies can be put in place to improve the surveillance and response to HIT-related adverse events at organizational, regional, national and the international levels?

The goal of this discussion will be to transfer learnings and to identify gaps and areas for further work. The audience will also be given an opportunity to consult and seek advice from panelists about practical issues in preventing and managing large-scale events.

**Statement of the Panel Organizer**

All panelists have agreed to participate. The panel is supported by the International Medical Informatics Association (IMIA) Working Group on Technology Assessment and Quality Development.

**References**

2. Emerson K. Fiona Stanley Hospital systems crash. The Western Australian. 2015 Feb 16.
Data Quality in Clinical Data Research Networks (CDRNs)

Panelists
Allison B. McCoy, PhD\(^1\), Michael G. Kahn, MD, PhD\(^2\),
Lemuel R. Waitman, PhD\(^3\), Jason N. Doctor, PhD\(^4\)

Moderator
Rahul Jain, MPH, CPHIMS\(^5\)

\(^1\)Tulane University, New Orleans, LA; \(^2\)University of Colorado, Denver, CO;
\(^3\)Kansas University Medical Center, Kansas City, KS; \(^4\)University of Southern California,
Los Angeles, CA; \(^5\)Louisiana Public Health Institute, New Orleans, LA

Abstract
National initiatives are increasingly leveraging the vast amounts of data availed by health information technology. PCORnet, the National Patient-Centered Clinical Research Network is one such example, and it aims to improve the ability to perform comparative effectiveness research by integrating data from clinical data research networks (CDRNs) and patient-powered research networks (PPRNs). One key component to the success of these networks will be ensuring the quality of the electronic health record and patient reported outcome data over time and across organizations. Clinical research informaticians have indicated challenges to data quality stemming from data integration, data capture, data exchange, and data standardization. Discussion items include the importance of data quality, a framework and plan for assessing data quality, the effects of health information system infrastructure on data quality, and data quality considerations for patient reported outcomes. Shared learnings from PCORnet stakeholders committed to the network’s excellence should help inform the national dialogue.

General Description
In the current healthcare environment, patients, healthcare providers, academics, insurers, and healthcare organizations are actively seeking the best clinical evidence to foster informed medical decision-making. To date, our capacity to provide real world evidence that can be broadly applied has been limited, in part, by relatively small and restricted datasets. PCORnet, the National Patient-Centered Clinical Research Network, has the potential to transform clinical research by providing the research infrastructure to allow investigators to conduct large comparative effectiveness studies using data from clinical practice in a large defined population. With a focus on patient-centered outcomes and research engagement, the results of these studies will provide important information to support evidence-based clinical care recommendations and to better explain multi-factorial contributors to health and illness, such as genetic, behavioral, social, and environmental factors across a spectrum of patients. PCORnet will integrate data from 11 Clinical Data Research Networks (CDRN) and 18 Patient-Powered Research Networks (PPRN). The CDRNs will each function as integrated research networks, including clinical data from multiple healthcare systems. Central to success of the CDRNs and PCORnet will be ensuring access, security, and quality of the data and appropriate linkage of data over time and across organizations. The need to ensure data quality is of particular importance given the tremendous amounts and the multiple sources of data to be included in the CDRNs; consideration must be given to the governance, evaluation, infrastructure, integration, and interoperability of data entered into the network to support data quality.

Anticipated Audience and Topic Importance
The efforts of PCORnet stakeholders committed to the development of a highly usable and effective infrastructure have the potential to affect patients, researchers, care providers, payers, employers and health systems. PCORnet’s large, national scale and recent timeline of funding make it especially important to all engaged in clinical research, in particular to clinical research informaticians with an interest or need in managing or utilizing vast amounts of clinical and patient reported data to derive new insight. The capacity to compare agreed upon indicators of data quality requires a coordination of approaches to evaluation among different organizations. The topic may also be of
interest to health systems administration and IT stakeholders as they consider the resources required to manage the embedding of best practices into operations activity and decision making. The topic provides an opportunity for other PCORnet and distributed research network or clinical research stakeholders within the audience to share in their learnings to ultimately validate or inform the practices for assuring the data quality practices and processes required to ensure confidence in new insight.

Discussion Questions

1. Why is it important to assess data quality in CDRNs?
2. How can my site assess the quality of data in a CDRN?
3. How can infrastructure choices affect data quality?
4. What special considerations are there in assessing the quality of patient reported outcomes?

Brief Description of Panelists and Presentations

Allison B. McCoy, PhD is an Assistant Professor in the Department of Biostatistics and Bioinformatics at Tulane University School of Public Health and Tropical Medicine and a research investigator at the Ochsner Health System Center for Applied Health Services Research. She is also a co-investigator for the Louisiana Clinical Data Research Network.

Dr. McCoy will present an introduction to data quality in CDRNs and provide justification for data quality assessment. She will also discuss approaches to governance to ensure data quality in CDRNs, including environmental scans and data characterization reports.

Michael Kahn, MD, PhD is the Director of Research Informatics at Children’s Hospital Colorado and the Biomedical Informatics Core Director for the Colorado Clinical and Translational Sciences Institute (CTSA). He serves as the informatics co-lead in PEDSnet CDRN, an informatics core member in the PORTAL network, and the site informatics lead for the NCATS-funded ACT network.

Dr. Kahn will present the value of developing a comprehensive data quality assessment framework and plan for evaluating data quality across multi-institutional CDRNs, including work that has established a broad-based collaborative network of individuals interested in establishing common approaches, models, and methods that can be shared across data networks that would support comparative and comprehensive data quality metrics, and work that evaluates existing, well-established data quality tools against this evolving harmonized model for data quality assessment and data quality reporting.

Lemuel R. Waitman, PhD is the Assistance Vice Chancellor for Enterprise Analytics, Director of Medical Informatics, and Associate Professor in the Departments of Biostatistics and Internal Medicine at the University of Kansas Medical Center. He serves as principal investigator for the Greater Plains Collaborative CDRN.

Dr. Waitman will present an overview of the infrastructure for CDRNs with respect to data quality, including the informatics interdependencies between electronic health record systems and enterprise data warehouses, and the challenges associated with integrating multiple sources of data (e.g., claims information, REDCap) into a single data mart.

Jason N. Doctor, PhD is an Associate Professor in the School of Pharmacy and the Director of Health Informatics at the Leonard D. Schaeffer Center for Health Policy and Economics at the University of Southern California. He is a co-investigator for pSCANNER CDRN.

Dr. Doctor will present the unique considerations and opportunities for evaluating the quality of patient reported outcomes, including those associated with measures, standardization and interoperability. Dr. Doctor will also report on the unique needs for further investments in characterization, and the potentials for examining relationships between clinical data and patient reported outcomes.

Rahul Jain, MPH, CPHIMS is the Digital Health Analytics Manager at the Louisiana Public Health Institute and the Digital Health Director for the Louisiana Clinical Data Research Network, responsible for quantified-self device data, data quality, data characterization and data standardization. Mr. Jain also manages the Analytics team for the
Greater New Orleans Health Information Exchange and is leading the building of a transitions of care infrastructure with Orleans Parish Prison.

Mr. Jain will serve as moderator for the panelist presentations and discussion.

All panelists named above have agreed to take part on the panel.
Panel: Public Implementation Resources for Genomic Medicine

Josh F Peterson, MD, MPH1; Marc S. Williams, MD2; Casey Overby, PhD2,3; Robert R. Freimuth, PhD4; Iftikhar Kullo, MD, PhD1

1Vanderbilt University School of Medicine, Nashville, TN; 2Geisinger Health System, Danville, PA; 3University of Maryland School of Medicine, Baltimore, MD; 4Mayo Clinic, Rochester, MN

Abstract

Genomic Medicine programs have proliferated as the evidence for implementing compelling pharmacogenomics use cases or targeted therapy accumulates. However, current electronic health records (EHRs) are under-developed for manipulating genomic data or communicating results to patients and providers. Three NIH-funded networks – Electronic Medical Records and Genomics (eMERGE), Pharmacogenomics Research Network (PGRN), and Implementing GeNomics In PracTicE (IGNITE) -- are creating health information technology (HIT) resources to help fill these gaps. Members of these consortia, which span academic medical centers and integrated health systems, are working to create educational tools, patient engagement technologies, clinical decision support resources, and common data models to exchange structured genomic results and guidelines. During this panel, speakers will review 1-2 resources that are available publically, and discuss how to leverage these resources to implement a new precision medicine or translational program. A panelist will describe methods and resources to encode genomic data and knowledge so that is can be shared between EHR environments. Secondly, a repository for storing and indexing genomic clinical decision support artifacts will be presented as well as efforts to use Infobuttons to provide just-in-time genomic medicine education. Finally, a panelist will present engagement technologies which can assist with communicating genomic results directly to patients.

Panel Overview

The recent growth of precision medicine has been fueled by striking variability in drug response and therapeutic outcomes, explained, in part, by genomic variation. Yet translating compelling use cases of Genomic Medicine to clinical care has been slow despite ongoing efforts to share implementation methods and develop best practice guidelines. EHRs have been identified as a key component of large-scale discovery and implementation; this is related in part to the complexity and scale of genomic data and in part to the need to integrate genomic risks with other clinical factors. However, the ability of EHRs to manipulate, transmit and communicate genomic data is surprisingly immature -- inhibiting both discovery and implementation. Known barriers include genomic results that are recorded as images or in other computationally inaccessible formats, the lack of an accepted standard for the interoperable exchange of data, and insufficient conventions for clearly communicating and displaying genomic data electronically.

Members of the panel have extensive experience with the implementation of genomic medicine at their respective institutions, and each belongs to one or more of the national consortia working to develop methods and best practices: the Pharmacogenomics Research Network (PGRN), Electronic Medical records and Genomics (eMERGE) Network and Implementing GeNomics in PracTicE (IGNITE).
The objectives of this panel will be to:
- Describe public resources related to implementation of genomic medicine
- Illustrate how these resources can be leveraged for implementation
- Identify the primary lessons learned when using the resources to inform implementations
- Discuss national resources that will help establish best practices for implementation of genomic medicine

This panel will follow the recommended format of 4 brief presentations with time for questions and discussion with the audience. Dr. Kullo will serve as a moderator and will ensure the presentations are well coordinated and contribute to the overall theme.

Presentation 1 (Dr. Freimuth): Standards to Support Genomic Medicine: Enabling the Exchange of Data and Knowledge

A significant investment of resources is required to implement genomic medicine within existing clinical electronic infrastructure. Standards that enable the exchange of genomic data and knowledge can reduce barriers to implementation. In this presentation, Dr. Freimuth will discuss efforts by HL7, ClinGen, and the Institute of Medicine Action Collaborative to develop robust methods for exchanging clinical genomic data. He will also describe projects by the PGRN Translational Pharmacogenomics Program (TPP), Clinical Pharmacogenomics Implementation Consortium (CPIC), and ONC Standards and Interoperability Framework (Health eDecisions) designed to facilitate the sharing of knowledge, including that expressed in clinical genomic guidelines and decision support rules, through structured knowledge representation schemes.

Presentation 2 (Dr. Peterson): Development and Initial Applications of a Repository for Genomic Clinical Decision Support Artifacts

Dr. Peterson will present the development and initial rollout of a repository for clinical decision support artifacts produced by institutions with genomic medicine programs. The repository reflects the myriad ways that institutions have developed clinical decision support and other customizations of their electronic health records (EHRs). He will discuss the functionality of the repository including capacity to store, index, and disseminate artifacts, and demonstrate their ability to act as a design model for new implementations. Policies for artifact submission, access, and site control over submitted content will also be discussed, as well as ways the artifacts are contributing to new translational programs. In addition to presenting the CDS Knowledgebase, Dr. Peterson will outline the process of transferring trusted guidelines into CDS interventions which are well integrated in clinical workflow.

Presentation 3 (Dr. Williams): Patient Powered Precision Medicine: Leveraging HIT to facilitate engagement

Patients are increasingly engaged in their health care through a variety of health information technologies. The role of the patient in genomic medicine remains uncertain given the volume, complexity and uncertainty of the information. However there are strong arguments supporting partnering with patients to manage genomic information. In this session, Dr. Williams will present the results of research in three areas: 1) the use of informatics tools to report genomic test results to both
patients and providers, 2) the development of a dynamic standards based interactive application, Genome Compass, that has the potential to enhance communication between patients and providers and 3) a prospective clinical trial that will study the impact of these tools on patient understanding, satisfaction, engagement and adherence as well as patient-provider communication. Dr. Williams will also address the HIPAA and privacy related issues related to the use of pharmacogenomic data within the EMR and patient-facing genomic reports.

Presentation 4 (Dr. Overby): Infobuttons as a mechanism for just-in-time education and clinical decision support for genomic medicine.

Integrating genomic information into the EHR requires just-in-time education of clinicians who may not have a background in the significance of particular genetic variants relevant for their patients’ care. Infobuttons are being explored as a decision support tool to provide such just-in-time education via context-specific links within the EHR. Dr. Overby will provide an update on progress to date and future plans to use infobuttons to link with external and institution-developed genomic medicine content within the EHRs of institutions pursuing genomic medicine projects.

Discussion Questions:
1) How do the resources discussed by the speakers work within vendor products and environments?
2) What are the current implementation resource gaps and are how could these to be addressed by federally funded projects?
3) What is the model for growing and maintaining genomic medicine HIT resources?
4) How do we assess the impact of genomic HIT tools on engagement of providers and patients?

About the Speakers:
The panelists were selected with broad knowledge of both applied informatics in the area of genomic medicine and direct involvement in the creation of many of the discussed resources. Robert R. Freimuth, Ph.D., is an Assistant Professor of Biomedical Informatics in the Department of Health Sciences Research, Mayo Clinic, and among many positions within the genomic medicine community, leads the PGRN Pharmacogenomics Ontology network resource and is past Chair of the AMIA Genomics and Translational Bioinformatics working group. Josh F Peterson, MD, MPH is an Assistant Professor of Biomedical Informatics and Medicine at Vanderbilt University Medical Center working with the IGNITE, eMERGE, and TPP consortia on implementation projects. Marc S. Williams, MD, FAAP, FACMG is a clinical geneticist, board-certified in clinical informatics and directs the Genomic Medicine Institute of the Geisenger Health System. Casey L. Overby, Ph.D. is a faculty member in the Program for Personalized and Genomic Medicine within the University of Maryland School of Medicine and Adjunct Investigator at Geisenger Health System involved in ClinGen, eMERGE, and IGNITE. The moderator, Iftikhar Kullo, MD is a Professor of Medicine at the Mayo Clinic with a focus on electronic health record-based genomic discovery and implementation of genomic medicine.

Affirmation
All proposed panel members have been personally contacted by Josh Peterson and have agreed to participate in this panel. The speakers report no conflicts of interest.
Title: Perioperative Clinical Decision Support: Improving Care of the Surgical Patient through Informatics

Moderator: Karl A. Poterack, MD
Assistant Professor of Anesthesiology
Diplomate in Clinical Informatics, American Board of Preventative Medicine
Medical Director, Applied Clinical Informatics,
Office of Information and Knowledge Management
Mayo Clinic
Phoenix, AZ

Panelist 1: Patrick Guffey, MD
Assistant Professor, Anesthesiology
Children’s Hospital Colorado
Aurora, CO

Panelist 2: Brian W. Pickering, MB, BCh
Assistant Professor, Anesthesiology
Mayo Clinic
Rochester, MN

Panelist 3: Bala Nair, PhD
Associate Professor, Anesthesiology & Pain Medicine
University of Washington
Seattle, WA

Panelist 4: Richard H. Epstein, MD
Professor of Anesthesiology
Diplomate in Clinical Informatics, American Board of Preventative Medicine
Certified Professional in Healthcare Information Management Systems, HIMSS
Vice Chair, Anesthesia Information Systems
Sidney Kimmel Medical College at Thomas Jefferson University
Philadelphia, PA

Abstract
The perioperative period is a complex environment that can benefit significantly from the implementation of clinical decision support (CDS), given the large volume and velocity of data, the presence of many distractions, and increasing documentation requirements. Panel participants will describe projects they have successfully implemented that relate to four separate domains of perioperative CDS. Implementation of post hoc provider-specific feedback through the use of dashboards will be discussed in the context of improving the quality of patient care. Guidance of provider activities at the point of patient care in the operating room through integration of data from physiologic monitors, anesthesia machines, and the electronic health record and delivery of timely advice will be presented. A CDS system to facilitate compliance with the new SCIP 4 protocol related to control of blood glucose following cardiac surgery will be described. Translation of published research regarding Bayesian statistical predictions for time remaining during surgical cases to a real-time whiteboard to facilitate operating room management CDS will be shown. Panelists will discuss issues related to implementing CDS in the context of modern electronic health systems in which access to necessary data in near real-time has proven to be challenging.
**General Description of the Panel**

The panel comprises 4 individuals and a moderator, all of whom are directly involved in the development and implementation of perioperative Clinical Decision Support (CDS) systems, and several who have published extensively in the area. Following Greenes’ characterization of CDS as “the use of information and communication technologies to bring relevant knowledge to bear on the health care and well-being of a patient,” the panel will discuss four different aspects of CDS as applied to the perioperative period. These include individualized post hoc reporting to improve performance, real-time recommendations based on application of artificial intelligence to physiologic and patient monitoring data to guide clinical care, alerts and reminders to facilitate compliance with clinical protocols, and statistical forecasting as an aide to operating room management. A brief description of each presentation follows:

**Moderator** Karl A. Poterack, MD

**Presentation:** Overview of Perioperative Clinical Data support

**Summary:** The scope of clinical decision support during the preoperative, intraoperative, and postoperative periods will be briefly summarized.

**Panelist 1:** Patrick Guffey, MD

**Presentation:** Driving Reporting and Quality Improvement

**Summary:** Healthcare providers seldom have access to self-serve analytics on their performance. I will highlight best practices in how to design, implement, and refine provider dashboards across multiple domains. In addition, change management strategies to support access to this information will be described. Brief case studies in performance improvement by using this methodology will also be presented.

**Panelist 2:** Brian W. Pickering, MB, BCh

**Presentation:** Reducing the Risk of Perioperative Harm: A Case Study in the Development, Implementation and Sustainability of an Informatics-Driven Approach

**Summary:** Preventable harm has become an important quality improvement target. As patients travel through their perioperative journey they encounter multiple healthcare providers and handoffs from one care location to the next. There is a risk of information being lost during transitions with resultant delayed or omitted care plan elements. In this presentation, a case history of the development, deployment, impact and sustainability of informatics tools developed to support compliance with SCIP 4 glucose control are discussed.

**Panelist 3:** Bala Nair, PhD

**Presentation:** Development and Use of the Smart Anesthesia Manager (SAM): An AIMS Based Real-Time Decision Support Module

**Summary:** Development and application of a decision support module that analyzes anesthesia information in near real-time to detect on-going clinical issues will be described. Detected clinical issues are brought to the attention of the anesthesia provider immediately so that remedial steps can be taken quickly. Factors to be considered when designing and implementing of a real-time decision support module will be discussed.
Panelist 4: Richard H. Epstein, MD

Presentation: Real time Decision-Support for Operating Room Management on the Day of Surgery

Summary: An important aspect of operating room (OR) management on the day of surgery is to understand when ongoing cases are likely to be completed, since actual durations differ the scheduled times. The duration that a case has been underway can be used to predict the median time when the case will exit the OR, based on Bayesian statistics. The translation of the theoretical framework of this approach to its application on the OR electronic whiteboard will be described.

Timeliness of panel and anticipated audience

There is an increasing need for clinical decision support to assist healthcare providers working in the perioperative area with escalating tasks related to complying with regulatory and compliance requirements, adhering to care protocols, following clinical pathways, managing workflow, and processing the voluminous data that is presented from physiologic monitors. However, lack of simple and efficient access to real-time patient care data within the major enterprise EHR systems is a major barrier to the application of real-time CDS. This panel will be of interest to those involved in the development, implementation, and use of perioperative CDS, including anesthesiologists and certified registered nurse anesthetists, nurses who work in preoperative and postoperative care units and intensive care units, surgeons, and administrators charged with oversight of billing, compliance, and regulatory issues.

Discussion questions to enhance audience participation

1. What issues have you encountered in accessing real time data from your electronic health systems?
2. What challenges have you experienced implementing CDS in the perioperative area?
3. What unintended consequences of CDS have you experienced in your organizations?
4. What governance processes do you have in place to approve implementation of CDS?
5. How do you monitor the effectiveness of CDS?
6. What successes have you experienced in the application of perioperative CDS?
7. What failures have you experienced in the application of perioperative CDS?
8. What regulatory requirements have your organization considered when developing and implementing CDS?

Panel organizer attestation that all participants have agreed to take part on the panel

All panelists have agreed to participate. Abbreviated, NIH-type Biosketch CV’s have been provided for each.

Panel Sponsorship

This panel was suggested by Vitaly Herasevich, Chair of the AMIA Intensive Care Informatics Working Group, who was involved in the preparation and submission of this proposal. The panel is sponsored by the Working Group.
Harmonization of ICD-11 and SNOMED CT – Not just mapping!
Practical and Theoretical Lessons & Benefits to Users and Implementers

Alan L Rector, MD PhD, James Campbell MD, Bedirhan Ustun MD PD, Christopher G Chute MD DPH, Harold Solbrig MSc

1University of Manchester, UK, 2University of Nebraska Medical Center, Omaha, Nebraska, USA, 3World Health Organization, Geneva, Switzerland, 4Johns Hopkins University, Baltimore, Maryland, USA, 5Mayo Foundation, Rochester, Minnesota, USA

Abstract

ICD and SNOMED CT were designed for different purposes – ICD for statistical reporting and epidemiological studies; SNOMED CT for documenting clinical care. They conform to different semantics – ICD is a closed monohierarchy in which sibling entities must be mutually exclusive and jointly exhaustive. SNOMED is an open polyhierarchy based on description logic. ICD has a strong requirement to focus on identifying diagnosis and underlying cause of death; SNOMED has an equally strong requirement to reflect the evolution of clinical understanding. ICD’s usage has clearly outgrown its historical structure. Updates occur at long intervals, and migration between versions is costly. To meet these problems and harmonize with SNOMED, ICD-11 uses a radically new three-component architecture: A “Common Ontology” that will correspond to an agreed subset of SNOMED, a “Foundation Component” structured around this ontology, and “Linearizations” that correspond to the existing classification, plus algorithms that link these component. The result is expected to simplify generating maps between SNOMED and ICD and to provide a stable platform for the evolution of ICD. The development has required clarifying the semantics of both ICD and SNOMED and has important implications for anyone working with either or attempting to harmonize other systems with them.

General Description

The panel will present and stimulate discussion on the underpinnings and practical outcomes of the architecture for ICD 11 and the harmonization of ICD-11 and SNOMED CT. Achieving the harmonization has required clarifying the semantics of both ICD and SNOMED and developing systematic means of using one to support the other. This is not a matter of “mapping”. The two systems and their requirements are too different. Rather it requires fundamental clarifications of their semantics, linkages between their conceptual models, and creation of a novel architecture for implementation. It has involved overcoming fundamental misconceptions and misunderstandings on each side: e.g. the distinctions between “ontological” definitions and “clinical descriptions”, between OWL/DL queries and concept expressions, and between “condition oriented” and “situation oriented” interpretations of the entities to be classified plus a clearer understanding of ICD’s “residual categories” – e.g. “not otherwise specified” and “not elsewhere classified”.

The panel will discuss these and other issues with respect to both in ICD-11 and SNOMED. More broadly, it will discuss the respective roles of logic based “ontologies” and the “classifications.”

It will also discuss the practical implications for users of both SNOMED and ICD-11, the further evolution of the ICD, and issues in migration from ICD-10. The architecture represents a means of reconciling the conflicting requirements for stability and support for change that bear particularly strongly on ICD. It is expected that in future ICD-11 will be more flexible and evolve more smoothly than previous versions. Using ICD and SNOMED together should become easier and more standardized, and the effort devoted to developing “mappings” should be reduced.

The issues raised are at the core of the Health Informatics agenda for achieving interoperability between clinical and epidemiological use cases. The existence of multiple conflicting terminologies continues to be a major challenge, often aggravated by advocates of one misunderstanding the purpose and requirements of another. The cost of migrating between versions or between terminologies has proved daunting. Although there have been many attempts at “mappings”, this is the first thoroughgoing attempt to harmonize two disparate systems that addresses the fundamental differences between them in both requirements and implementation and respects the semantics of each.
Presentations and Presenters

Goals and Constraints on the ICD: Bedirhan Ustun (WHO)

The ICD is an international standard that serves many purposes. It was developed originally for reporting mortality and morbidity. It has over a century of history behind it and a large body of expertise around the world in its use. However, many groups wish to use it far beyond the intent of original statistical categorization including for recording diagnoses in clinical and epidemiological studies, quality and safety reporting, and development of diagnostic groupings for outcomes and reimbursement. Variations in terminological, linguistic, and clinical practice across countries and over time add more complexity in comparing data. The comparability of various versions and meaningful exchange of data between different uses is a major challenge. It is therefore essential to move to an electronic infrastructure that will:

1. Enable users to ensure comparability of data captured in different settings by pointing to international reference standards for diagnostic categories (e.g. defining hypertension, anemia, depression etc.) independent of language artefacts.

2. Link the different diagnostic formulations used in different levels of health care in a consistent way so that primary care, secondary care and other health workers can exchange information reliably.

3. Link to standard terminologies (such as SNOMED-CT, LOINC, etc), Classifications (ICF, ICHI, ATC) and Ontologies (Gene Ontology etc) so that the electronic processing of data is performed in a standard way.

Overview and motivation: CG Chute (Chair: ICD Revision Steering Group)

ICD-11 aims to reconcile the conflicting requirements for greater flexibility and for continuing stability within ICD. At the same time it aims to achieve a harmonization with SNOMED CT. Since SNOMED CT is aimed at clinical practice and ICD primarily at statistical returns, many systems will need both. If the goal of collecting statistical data automatically from EHRs is to be achieved, then we need to be able transform SNOMED CT encoded data to ICD. Although maps exist for ICD-10, their production has been a long and expensive process fraught with problems and without any common principled foundation.

There are two fundamental ideas that underlie ICD-11 and its harmonization with SNOMED CT:

1. To separate the components that correspond to the existing and previous ICD-10 versions – the “Linearizations” – from a generic, polyhierarchical “Foundation Component.” The “Joint Mortality and Morbidity Linearization,” will be the direct successor to ICD-10, but alternative organizations are possible, e.g. for rare diseases, primary care, oncology, etc. It is hoped that this same mechanism will allow for representation of ICD 10 and existing clinical modifications, but this has yet to be demonstrated.

2. To use a “Common Ontology”, agreed jointly with SNOMED CT, and sharing its description logic based semantics, as the skeleton for the “Foundation Component”. The “Foundation Component” contains much more than the “Common Ontology”, but building it around the “Common Ontology” is expected allow largely algorithmic mapping between the SNOMED CT and ICD-11 and agreement on common text definitions.

Ontologies and Classifications, Definitions and Descriptions: The Role of a “Common Ontology”: Alan Rector (ICD revision steering group)

“Ontologies” and “Classifications” in the senses used by SNOMED CT and ICD, respectively, are fundamentally different. Ontologies are polyhierarchical; classifications are mono-hierarchical. Ontologies use logical semantics based on meaning; Classifications use pragmatic semantics based on use. Likewise “definitions” corresponding to the ontological meaning expressed in logic are fundamentally different from clinical descriptions. For example, ontologically, a “myocardial infarct” is (roughly) “death of some part of the myocardium due to ischemia”; clinically it can be described as “Manifesting variously as sudden death, acute crushing chest pain,…”. These two key distinctions and their effect on the ICD SNOMED CT harmonization process will be discussed as will the problem of choosing amongst the possible alternative semantic interpretations of SNOMED CT and ICD definitions’.

Implementation issues: Evaluating the Completeness and Correctness of the Classification/Ontology Linkage: Harold Solbrig (SNOMED CT/IHTSDO)

The foundation ICD-11 classification system is defined using a set equivalence expressions, either directly with existing SNOMED CT concepts or indirectly, using “post-coordinated” expressions to identify the intended meaning of the category. These equivalence expressions must determine the conditions that are both necessary and sufficient
for a given clinical event to be recorded as belonging to a particular category. This step, in turn, drives the linearity process, where the specific ICD-11 categories are combined using inclusion and exclusions that determine where a specific event should be categorized when more than one category is possible.

It is not possible to do this second step directly, because the requirement that the linearizations be single hierarchies means that, if a category is logically a subclass of more than one class, then one parent class must be chosen and the others excluded. This cannot be done within the Common Ontology itself because to do so would violate its logic-based semantics. Instead, it must be done by querying the resulting structure of the Common Ontology using a query language. These issues will be explained and the query language being developed within the IHTSDO presented briefly. Current practical efforts are being proven in the cardiovascular/circulatory sections.

**Practical experience from Mappings and semantic alignment: Jim Campbell (SNOMED CT/IHTSDO)**

Recent publication by the IHTSDO of rule-based mapping from SNOMED CT to ICD-10 has revisited and quantified differences in conceptual models employed by the two terminologies. Observations regarding semantic mismatch in ICD-10 maps are informing the alignment between SNOMED CT and ICD-11. Co-development requires that complete ontological definitions be developed for ICD-11 classes and that these be fully defined in expressions employing the common ontology subset of SNOMED CT. Exemplar cases of semantic misalignment from ICD maps will be presented, relating this to the semantic analysis and redefinition of ontologies that has developed from harmonization case studies. Preliminary results for the cardiovascular chapter indicate that this alignment is non-trivial. The issues raised by this experience will be discussed.

**The Future Directions and other Features of ICD-11: CG Chute (Chair: ICD revision steering group)**

The primary efforts of the present revision remains to deliver to the global community a disease and condition classification that is scientifically current, broadly useable, and addresses the needs of many stakeholders. The current architecture displays sufficient flexibility to accommodate a large spectrum of use cases through multiple linearizations, each anchored to the semantic core of the Foundation Component, and in turn to SNOMED CT.

Although not yet fully populated by the time of its initial release, ICD-11 will eventually include a rich and robust Content Model in the Foundation Component. The Content Model will incorporate structures for diagnostic criteria (to accommodate multiple paths and technology settings), more complete etiologic and genomic detail, a fully coded rendering of disease and condition characteristics such as anatomy, disease specific severity, impact on human functioning, and possibly additional resources. These will be refined after the official ICD-11 release, to contribute to the scientific completeness and usefulness of the ICD-11 information family into the future.

**Questions for Discussion**

- What might be the migration path to ICD-11, with or without SNOMED CT?
- Aren’t the Linearizations enough? Do implementers have to deal with anything else? Do clinicians?
- Why isn’t SNOMED CT enough? Why isn’t ICD enough? Why both? Does this mean we can just use ICD? Just use SNOMED CT?
- How will these developments affect coders? Clinicians? Implementers?
- Will you have to understand Description Logics to use ICD-11? What benefits if you do? Will you have to understand SNOMED CT to use ICD-11? ICD-11 to use SNOMED CT?
- What will be the best way to manage proprietary terminologies in conjunction with ICD-11?
- Will CM versions of ICD still be needed?

**References**

Developing Natural Language Processing Systems for Healthcare

Panelists
Ruth Reeves, PhD¹, Wendy W. Chapman, PhD², Dezon Finch, PhD³, and Jennifer H. Garvin, MBA, PhD⁴

Organizer & Moderator
Glenn T. Gobbel, DVM, PhD, MS¹,⁵

¹Tennessee Valley Health System VA, Nashville, TN; ²University of Utah, Salt Lake City, UT; ³Tampa VA, Tamp, FL; ⁴Salt Lake City VA Medical Center, Salt Lake City, UT; ⁵Vanderbilt University, Nashville, TN;

Abstract
Natural language processing (NLP) systems can help uncover medical knowledge buried within clinical free text. Despite multiple demonstrations of the clinical utility of NLP, its use for biomedical research and patient care remains limited. A detailed overview of the steps involved and potential challenges associated with creating an NLP system could help potential users of such systems; they could use this knowledge to assess feasibility and identify the types of expertise required for system development and implementation. This panel will delineate four key tasks commonly required when developing a clinical NLP system: 1) document selection; 2) annotation; 3) NLP software development; and 4) clinical implementation. Presenters will explain the role of each task in NLP system development, define the subtasks required, describe approaches for estimating the effort and time requirements, and delineate approaches and tools for assisting with task completion. Attendees will acquire an understanding of the common requirements for NLP system development, an appreciation of potential tools to help with the process, and be better able to assess whether and how to approach using NLP in their own biomedical centers.

General Description
NLP identifies and extracts information from unstructured free text and converts it into a structured form for further use by medical researchers and providers. Multiple studies have demonstrated that this method of data extraction can effectively work with both clinical research and care, which has led to an increased interest in employing this approach to address other medical issues. Selected examples of successful demonstrations include extraction of medication for visualization and reconciliation,¹ screening for patients suffering from traumatic injuries,² identification of post-operative complications,³ detection of signs of infection and cases of influenza for use in biosurveillance,⁴,⁵ quantitation of the use of evidence-based psychotherapy for treating post-traumatic stress disorder patients within the Department of Veterans Affairs,⁶ extraction of clinical decision support information from research reports included within the PubMed database,⁷ detection of patients for colorectal cancer screening,⁸,⁹ and determination of the efficacy of using influenza vaccines to prevent pneumonia.¹⁰ Despite evidence demonstrating the potential benefits of using NLP, medical personnel have, to date, rarely employed this approach to address medical problems or improve patient care.¹¹ Researchers and clinicians interested in using NLP may be hesitant to advocate for its implementation without having the knowledge needed to assess utility and feasibility. The goals of this panel will be to provide that knowledge and to encourage a dialogue between NLP system developers and potential users in order to identify and address barriers to NLP system use. The panel will consist of 4 panel members and a moderator, all of whom have direct experience with the process of developing or implementing medical NLP systems. Each panelist will focus on one of the 4 common steps in NLP system development, 1) document selection, 2) annotation, 3) NLP software development and adaptation, or 4) clinical implementation; panel members will describe lessons learned and elucidate the considerations for accomplishing each step and the challenges that may arise. They will also describe and provide handouts with web page links to current and upcoming tools that may assist with the task. By attending these presentations, potential users will gain an understanding of the overall NLP system development workflow. They will also be able to better assess the applicability of NLP to their own clinical problems of interest.

Structure & Discussion Topics
The moderator will first provide a 15 min overview of the motivation for the panel followed by brief descriptions of the 4 key tasks commonly required for developing an NLP system for use in healthcare. Panelists will then spend 15 min focusing on one of the 4 tasks, and presentations will generally include descriptions of the role of the task in NLP system development, panelists’ own experiences including workflows and lessons learned, tools implemented or used, and barriers to task completion. The last 30 min of the presentation will be for questions from the audience and for discussion between the panelists and audience. Topics for discussion include but will not be limited to:
Capabilities of NLP for medical research and care: Potential users need to be aware of both opportunities that NLP provides and also its current limitations. Specific discussion questions include: 1) What are the clinical and research questions that NLP can address? 2) What biomedical research and clinical questions is NLP currently unable to address? 3) What clinical or research questions are best addressed using approaches other than NLP? 4) What do potential and current users view as the greatest barriers to NLP use?

Development and implementation issues: Currently, there are no “turn-key” NLP applications available for addressing non-trivial medical issues. Developing and implementing an NLP system is often a team effort requiring extensive knowledge in computer science, linguistics, information technology, and implementation science in addition to clinical expertise. Given this current state of clinical NLP, questions to be addressed include: 1) How are the time and expertise requirements for developing and implementing a NLP system estimated? 2) What are the requirements and challenges for completing each of the four steps typically required when developing an NLP system? 3) What do current and potential users see as the major barriers to developing NLP systems for healthcare? 4) Will there ever be “user-friendly” NLP tools developed for use by non-experts?

NLP opportunities: The increase in computer processing speed and infrastructure over the past 10-20 years has increased the feasibility of using NLP to address medical questions, but there appears to be substantial room for growth in its use. Multiple clinical issues that might be helped by the use of NLP represent opportunities for investigation but remain unaddressed. Furthermore, the nature of medical free text may present unique NLP challenges. Questions for discussion with respect to clinical opportunities and NLP challenges include: 1) What NLP challenges are particular to healthcare? 2) What opportunities are there for extending the utility of NLP in healthcare? 3) What clinical questions might particularly benefit from the use of NLP?

Anticipated Audience & Importance of the Topic
This panel should be of particular interest to 1) those intrigued by the promising results generated using NLP but who have a limited understanding of its potential and limitations as well as the full system development process, 2) those interested in tools and workflows that can assist with the development of an NLP system for use in health research or clinical care, and 3) current NLP system users and developers who can share their own experience with barriers to use and would like to learn about the hurdles experienced by potential NLP users in the audience. This topic is timely in that we appear to be at a tipping point with respect to expanding the utility of NLP; a key barrier to its use is not so much its ability to extract useful information but rather getting the techniques and tools into the hands of potential users. Each of the panelists has experience developing and implementing NLP systems for medical research and clinical care. They will share their experiences and describe the tools and workflows they have used in the process, which will provide audience members with the knowledge needed to assess the feasibility and requirements for NLP system development and implementation.

Brief Description of Panelists & Presentations
Dezon Finch, PhD is a health science specialist at the HSR&D/RR&D Center of Innovation in Disability and Rehabilitation Research VISN 8 (CINDRR). Dr. Finch’s research interests include use of statistical approaches to text-mining of medical datasets and ontology generation, using machine learning to detect semi-structured text elements and sections in medical progress notes, and using information extraction for evaluating quality of care. He has served as an investigator within the VA Consortium on Health Informatics Research (CHIR) developing ontologies and schemas for signs and symptoms of post-traumatic stress disorder. He is currently a co-investigator on two VA Investigator Initiated Research (IIR) projects that use NLP; one to determine risk factors for the development of pressure ulcers following spinal cord injury and another to evaluate the quality of pain reassessments in primary care settings. He was PI on a recently completed study that used NLP to evaluate adherence to treatment protocols for veterans with TBI symptoms. His presentation will focus on the issues to consider when selecting clinical documents for developing an NLP system and describe his own experiences and approaches used in completing this task.

Ruth Reeves, PhD Ruth Reeves, PhD is a health services research scientist at the Tennessee Valley Health System VA and an assistant professor in the Department of Biomedical Informatics at Vanderbilt University. She is currently principal investigator on a nationally funded merit award which deploys NLP methodology to temporal reasoning within the electronic health records of veterans suffering from PTSD and is a co-investigator on six nationally-funded projects focusing on the development of NLP systems for clinical use. Dr. Reeves will present and facilitate discussion of the tasks involved in setting up an annotation task and successfully obtaining high quality annotations that meet the needs of the NLP development process. This includes matching annotator domain of expertise with the subject matter of information extraction, annotator training, best practices for resolving annotation disagreements, prioritizing annotation tasks in terms of anticipated information consumption. Her presentation will go thru a typical workflow cycle for an annotation project to illustrate these and other operational considerations.
**Wendy W. Chapman, PhD** is the chair of the University of Utah Department of Biomedical Informatics. Her research focuses on developing and disseminating resources for modeling and understanding information described in narrative clinical reports. She is interested not only in better algorithms for extracting information out of clinical text through NLP but also in generating resources for improving the NLP development process (such as shareable annotations and open source toolkits) and in developing user applications to help non-NLP experts apply NLP in informatics-based tasks like clinical research and decision support. In this panel, Dr. Chapman will describe the steps required in creating an NLP tool for a targeted task. With a real-life use case, she will discuss the types of tools and resources that are available to create or customize an NLP tool, including knowledge resources, NLP pipelines and tools, and evaluation tools.

**Jennifer H. Garvin, MBA, PhD** is a research health scientist within the VA Salt Lake City Health Care System and associate professor at the University of Utah. Dr. Garvin’s research interests include automated quality measurement, stakeholder engagement for requirements gathering and applied uses of NLP. She has professional training and research experience in terminologies, classifications, healthcare standards, implementation science methods and NLP. Dr. Garvin is currently the principal investigator on a project that will use NLP to improve post-CHF discharge beta blocker titration, according to the guideline-recommended care, and thereby reduce hospital readmissions and improve patient outcomes, and she is a co-investigator on three additional nationally-funded research projects involving the clinical use of NLP for information extraction. Dr. Garvin will share her experience and the workflow she has used in developing and implementing NLP systems for clinical use.

**Glenn T. Gobbel, DVM, PhD, MS** will moderate the panel, and he is a research assistant professor in Health Services Research at the Tennessee Valley Health System and within the Department of Biomedical Informatics at Vanderbilt University. His current research interests are the generation of algorithms that improve the performance and enhance the clinical utility as well as development of applications to assist in the development of NLP system. His work has focused on the development of near-real time NLP tools for extraction of medical information from free text and machine-learning based tools that rely on online learning for assisted annotation. He is currently a co-investigator on six nationally-funded projects focusing on the development of NLP systems for clinical use.

**Participation Statement:** All proposed panelists have agreed to participate in the panel.

**References**

Didactic Panel
Natural Language Processing for Phenotype Extraction: Challenges in Extraction and Representation

Guergana Savova, PhD¹, Rebecca Jacobson, MD, MS², Joshua Denny, MD, MS³, Nicole Washington, PhD⁴, Harry Hochheiser, PhD²

¹Children’s Hospital Boston, Boston, MA; ²University of Pittsburgh, Pittsburgh, PA; ³Vanderbilt University, Nashville, TN; ⁴Lawrence Berkeley National Labs, Berkeley, CA

Abstract

Previous work has demonstrated the benefits of using Natural Language Processing (NLP) techniques for extracting diagnoses from clinical texts and assigning patients to categories for cohort identification and GWAS studies using electronic health record (EHR) data. However, there is a growing awareness that understanding cancers and other complex or rare diseases will require richer phenotypic models that describe the development, progression, and location of specific signs and symptoms. Computing over these “deep phenotypes” will require NLP methods that can interpret individual documents, summarization approaches that can combine information from multiple notes and document types (e.g., radiology, pathology, and clinical notes) to build a longitudinal patient, and structured data models that will facilitate the integration between NLP and existing discrete data and comparison of these phenotypic models. This panel will discuss challenges inherent in developing both data models and the NLP tools. Strategies involving the use of NLP pipelines (cTAKES), clinical data models (FHIR), and phenotype data exchange formats will be discussed. Attendees will gain an understanding of the NLP challenges in extracting deep phenotype information, strategies for representing phenotype data, and emerging systems and tools that are applying these techniques to challenging translational problems.

Description

This panel will provide multiple perspectives on the challenges of using Natural Language Processing to extract detailed phenotype models from clinical texts. Presentations will cover technical issues including NLP pipelines and phenotypic models, experience in application to specific diseases, and translational linkage to genomics. This breadth of discussion will provide participants with an understanding of the current landscape and an appreciation for the potential benefits of combining NLP with principled phenotype modeling:

• Geurgana Savova, project lead for the widely used cTAKES NLP pipeline, will discuss new developments and current efforts in extending the capabilities of NLP tools.
• Rebecca Jacobson will present an overview of the overall landscape and perspectives from the application of NLP and phenotyping to cancer informatics.
• Joshua Denny will describe lessons learned from eMERGE and other large-scale phenotyping projects integrating NLP with other data across the EHR.
• Nicole Washington will introduce translational applications for structured phenotype descriptions.
• Harry Hochheiser will discuss issues involved in data models for phenotype representation.

Need

• Clinical NLP continues to grow in importance
• Characterizations based on simple categorization are not sufficiently nuanced for translational research of complex diseases
• Models of computing over complex phenotypic descriptions are nascent and barely emerging. Specifically:
  • Although temporal models exist, relatively little work has been done in using NLP to interpret complex descriptions of patient state over time.
o Building longitudinal models of patient histories requires summarization logic not found in many NLP pipelines
o Construction of high-level overall patient phenotypes models that can be exchanged and integrated across contexts
o Although integration with clinical data models might be beneficial, models such as FHIR are not yet widely adopted.

Clinical and translational researchers working in these areas currently face significant challenges in identifying and evaluating appropriate tools and data models. This panel will provide a variety of perspectives that will help participants better understand how to engage in NLP-based phenotyping efforts.

**Audience**
- Clinical researchers interested in extracting structured phenotypes from EMR data.
- NLP developers working on new techniques for addressing challenges in interpreting clinical text.
- Participants in PCORNet or other efforts aimed at aligning phenotype data from multiple diverse sources.
- Translational researchers searching for model systems for human diseases.

**Discussion Questions**
- What level of detail is required in NLP, summarization, and exchange models of phenotypes?
- How can models be effectively validated?
- How does data extracted via NLP compare with structured data for phenotype sensitivity, precision, and granularity?
- How do variations in care across institutions/contexts impact the generalizability of these models?
- How can summarization logic be expressed in some computable form?
- What are the advantages and disadvantages of using OWL representations? FHIR?
- How can the tradeoffs between consistency of models and flexibility for adaptation be best balanced?

**Participants’ Assent**

All participants have agreed to take part on the panel.
Needs of the Digital Native:

Adolescents and Access to PHRs

Catherine Arnott Smith, AMLS, MA, MSIS, PhD
Fabienne Bourgeois, MD, MPH
Pamela Charney, MS, MS, PhD

1University of Wisconsin-Madison, Madison, WI; 2Boston Children’s Hospital, Boston, MA; 3Bellevue College, Bellevue, WA

Moderator: Patricia Flatley Brennan, RN, PhD, FAAN, FACMI

Abstract

PHRs present particular challenges in pediatrics and the pediatric PHR is a particularly timely, while controversial, issue in medical informatics. This is true because minor patients present “unique health privacy and confidentiality standards,” different users within a family unit have different rights to access different information, and there are “extensive variations” in PHR implementation and access across the United States. This panel considers the challenges, benefits and risks of access to PHRs by adolescent patients. In order to address the needs of this population, we have to support access policies that handle age-specific changes in PHR ownership, as well as encourage EHR functionality that allows appropriate identification and classification of confidential information. We also offer two examples of how these challenges play out in context of adolescents’ lived experiences. How is care coordinated for adolescents with chronic health conditions when PHRs are not available? Can PHRs play a role in assisting adolescents and young adults living with chronic illness to develop self-advocacy and self-awareness?

Barriers to PHR Access

Personal health records (PHRs) have been defined as “collections of health or wellness data arising from multiple sources about an individual’s health … managed, controlled or shared by that individual or designate.” These can be paper-based or electronic systems. PHRs’ information management functions include reporting immunization data, laboratory test results, visit summaries, and problem lists. They have received increased attention from biomedical informatics as increasing amounts of health data become available in EHRs to which PHRs can be tethered. For this reason, PHR growth parallels that of EHRs.

The potential of PHRs for recordkeeping by families, beyond the individual, has been formally investigated since 1964. In the 21st century, use of PHRs in pediatrics presents a complicated situation for communication and information-sharing alike. First, pediatrics has “lagged behind other specialties” in using electronic tools, and few PHR systems have been developed. Second, in pediatric practice, patients occur in dyads and triads. Parents are the first responders, primary communicators, and information managers for their children. Bourgeois, Taylor, Emans, Nigrin & Mandl have written of the challenges these facts of life pose for PHR development: minors present “unique health privacy and confidentiality standards,” different users have different rights to access different information, and there are “extensive variations” across states and localities in the US.

Panel Description

This panel begins with presentation of the challenges, benefits and risks of access to PHRs by adolescent patients. We then present two examples of how these challenges play out in context of lived experiences: first, the management of chronic illness without a PHR; second, the acquisition of self-advocacy skills for negotiating academic accommodations in the high school and future college classroom.

Patricia Flatley Brennan, RN, PhD, will serve as moderator for the panel. Dr Brennan is the Lillian L. Moehlman Bascom Professor, School of Nursing and College of Engineering, University of Wisconsin-Madison, Madison, Wisconsin. She developed the ComputerLink, an electronic network designed to reduce isolation and improve self-care among home care patients and directed HeartCare, a WWW-based tailored information and communication
service that helped home-dwelling cardiac patients recover faster, and with fewer symptoms. Dr. Brennan directed Project HealthDesign, a RWJ-funded initiative designed to stimulate the next generation of personal health records. Brennan leads the Living Environments Laboratory at the Wisconsin Institutes for Discovery, which includes a 6-sided virtual reality CAVE that her group uses to re-create visually every environment on earth, and develop new ways for effective visualization of high dimensional data. Supported by AHRQ, her vizHOME group explores the impact of household contexts on personal health information management.

Fabienne Bourgeois is a Pediatric Hospitalist and Clinical Informaticist at Boston Children's Hospital. As the Medical Director of Patient-Facing applications, she has helped develop and implement a patient portal specifically designed to meet the needs of a pediatric and adolescent population.

Pamela Charney is a former practicing registered dietitian specializing in pediatrics and clinical care. Her previous research focused on nutrition-related issues ranging from development of a pediatric nutrition center of excellence to evaluating nutrition status of underserved groups who receive home care. She is currently collaborating with the Informatics Consortium at Bellevue College as part of their $12 million DoL project and has another project in development investigating health outcomes associated with technology use by adolescents with chronic health conditions and their caregivers.

Catherine Arnott Smith is an Associate Professor at the School of Library & Information Studies, University of Wisconsin-Madison. Her research interests and publications center on consumer interactions with clinical information systems through text, particularly consumer health vocabulary, and in domains ranging from electronic medical records to public libraries.

All participants have agreed to take part in the panel.

BOURGEIOS

Adolescent patients rarely have full control of their personal health records (PHRs), instead relying on parents and guardians to share control. This shared access necessarily changes over time, as a result of developmental and age-appropriate considerations, as well as guardianship arrangements. In particular, once a patient reaches the age of majority (generally at age 18 years in most states), the patient becomes the sole guardian of the PHR. However, the transitioning adult’s readiness to assume this new role, in part, depends on her prior experience and access to a PHR.

Most current PHR ventures either limit access to the application or restrict the available information in the adolescent PHR due to the unique ethical and legal challenges experienced by this population. The greatest difficulty involves protecting the adolescent’s legal right to privacy and confidentiality within a shared access model. Many medical encounters with adolescents come with assurances that what they report will remain entirely confidential, so that personal health information pertaining to reproductive health, sexually transmitted illnesses, substance abuse and mental health will not be communicated to their parents without their consent. Similarly, parents may share information with the expectation of privacy and do not want the information shared with the adolescent. As it turns out, these types of confidential information are pervasive throughout most patients’ electronic health records (EHRs).

In order to address the needs of the adolescent population we have to support access policies that handle age-specific changes in PHR ownership, as well as encourage EHR functionality that allows appropriate identification and classification of confidential information.

This discussion will focus on the options for shared access models for adolescent PHRs (along with their respective benefits and risks), as well as handling of medical content in the EHR requiring special consideration to minimize inadvertent disclosures through the PHR.

CHARNEY

Use of pediatric personal health records (PHR) facilitates management of chronic illness in childhood. However, because of challenges associated with privacy and security of adolescent health information many healthcare organizations severely limit access to personal health records (PHR) for adolescent patients. Therefore, families caring for adolescents with chronic illness often lose access to the benefits of using their child’s PHR. Lack of access to the PHR has significant consequences including poor care coordination, duplication of tests, and poor communication.

Some healthcare organizations do allow parents limited access to their adolescent’s PHR, particularly for secure messaging with providers and appointment scheduling. When even limited access is not available, parents/caregivers of adolescents with chronic illness are forced to spend significant time coordinating care.
The situation is made even more complex when care is provided through more than one healthcare organization, each of which may have different policies regarding access to adolescent health information.

This presentation will focus on the challenges associated with coordinating care for adolescents with chronic health conditions when PHRs are not available and offer suggestions for improving access while maintaining appropriate privacy for the adolescent.

SMITH

PHRs seem to have a motivational effect on patients to serve as monitors of their own health data, from diabetes passports to Project Health Design. People living with chronic illnesses are more likely than others to maintain “summaries of their health histories, medications, and physician names” and this translates to a high interest in using PHRs—these people were found in one study to be 25% more likely to adopt these technologies.

Adolescents living with chronic illness are thus an important potential audience for PHR technologies. These heavy users of technology are transitioning into adulthood not only in the clinic, but in the classroom. The academic accommodation process for students with disabilities, like the Social Security Disability Insurance system described by Lytle-Kosola et al., uses health information as “a primary input to facilitate movement.”

The speaker agrees, with Lytle-Kosola et al., that consumer health IT can used to reduce the “information-gathering” burden for student clients. This is a population very likely to benefit from early familiarity with PHRs. When we shut out, or partially restrict, adolescent patients’ use of PHRs we create a knowledge burden and a resource challenge for the newly emerging adult they are engaged in becoming. This presentation will use examples from an ongoing research study to illustrate the role of clinical documentation in young adults’ self-advocacy and self-awareness around disability and academic accommodations.

Expected Discussion

Panel attendees will be encouraged to share their opinions and experiences. The goal is to discuss the feasibility of possible solutions to the challenges outlined in the panel presentations, and to contribute to ideas for a medical informatics research agenda moving forward.

References


Jonathan M. Teich, MD, PhD \textsuperscript{1,5}, Hamish Fraser, MBChB, MSC, \textsuperscript{1,2}, Eric Perakslis, PhD\textsuperscript{1}, Shefali Oza, MSc\textsuperscript{3}, Darius Jazayeri, MEng\textsuperscript{4}

\textsuperscript{1}Harvard University, Boston, USA. \textsuperscript{2}University of Leeds, Leeds, UK. \textsuperscript{3}London School of Hygiene and Tropical Medicine, London, UK. \textsuperscript{4}ThoughtWorks Uganda Ltd., Kampala, Uganda. \textsuperscript{5}Elsevier, Philadelphia, USA.

Abstract

In response to the large West African Ebola outbreak beginning in 2014, Ebola Treatment Centres (ETCs) were established to provide efficient care for suspected and confirmed Ebola patients while minimizing risk of cross infection to staff and other patients. Critical clinical information entry and review tasks such as assessments, ordering, and administration of drugs and fluids must be performed rapidly in highly challenging conditions because workers wear personal protective equipment (PPE) which reduces dexterity, vision and comfort. Additionally, it is vital to have information tools to conduct important functions, such as contact tracing to limit the spread of new cases and trials of new diagnostic/therapeutic approaches.

In this panel, members of two teams providing rapid development and implementation at different centres describe the unique situational needs and challenges, software and hardware considerations and solutions, user engagement throughout testing and rollout, and approaches to advanced functionality such as analytics and clinical decision support. The audience will be better able to plan future implementations in challenging environments; to include information management in advance planning for future crises; and should be able to leverage available technologies and strategies for effective design and deployment of crisis information systems.

Panel Description

The 2014 West African Ebola outbreak is unprecedented in its scale and spread, with over 24,000 suspected or confirmed cases and over 10,000 deaths through mid-March 2015. In response to the outbreak, Ebola Treatment Centres (ETCs) have been established; these specialized facilities must provide efficient care for suspected and confirmed Ebola patients while minimizing risk of transmission of disease to staff and other patients\textsuperscript{1}.

As with any medical facility, ETC operation involves information-intensive tasks, including registration and admission; frequent patient status assessment, recording and reviewing; ordering, dispensing and administration of medications, fluids, and blood products; ordering and reporting of a limited number of laboratory tests. Due to the focus on epidemic-scale infectious disease, tracing of contacts of each patient is also critical. Because optimal treatment for Ebola is not completely established and new treatments are under development, it is also important to capture information about status, treatments and outcomes in order to improve care in the future.

Many of these information tasks must be performed in the “red zone” where patients with suspected or confirmed Ebola are located. The red zone presents several challenges to all operations, including the difficulty of working in personal protective equipment (PPE) which significantly limits visual resolution and dexterity and also limits available time in the zone due to its discomfort. Since materials in the red zone are considered contaminated, they cannot be transferred out into the “green zone”; thus paper recording is not an option even if it were otherwise satisfactory. Moreover, electronic data recording is essential for proper documentation and communication, and especially for data collection, surveillance and research to improve care. The user interfaces of typical EMR applications and data collection tools, though, are difficult or impossible to use in the red zone environment. Innovative design and development is required to permit quick, usable, comprehensive information recording and review. In addition, innovative hardware and infrastructure design may be needed due to other environmental issues.
A number of organizations have risen to the challenge, rapidly developing and implementing Eb
l systems and tracking systems for ET
cs. One of these ET
cs was established at Kerry Town, outside of Freetown, 
Sierra Leone, under the auspices of Save the Children International; the development team included staff from 
Thoughtworks as well as several additional members of the OpenMRS community. At this centre, the team has 
deployed a novel interactive EMR with a wide range of administrative and clinical functions for green and red zone 
operation, using a tablet-based user interface suitable for PPE use with large text and touch targets, minimal typing, 
a high-contrast colour scheme, and functional design for rapid and easy operation with this interface. Other 
treatment centres have been established under the auspices of Médecins Sans Frontières; this team developed a 
system, also on an OpenMRS base, with assistance from a development team at Google Inc. It has placed 
additional emphasis on clinical trials management, mobile apps for contact tracing and epidemiology, and 
infrastructure advances such as resilient power, mesh networking and solvent-resistant casings.

A critical success factor in the deployment of these systems is working with the local healthcare team and leadership on workflow, implementation planning, and iterative system acceptance. The implementation leaders from the 
informatics teams had extensive discussions with all clinician roles, epidemiologists, programmers/developers, researchers, IT/connectivity staff, hygiene/infection control staff, and other organizations involved in the response. Such discussions were and need to be a continuous, rather than one-off, process given the dynamic nature of emergencies. In these conversations, important system requirements were conveyed, such as the need for 
waterproof tablet PCs and the real, most important everyday information elements. They also revealed what subset of functions constituted a minimum viable product (MVP) for the system, so that key modules could be rolled out 
rapidly and additional ones layered on as they were developed. Continuous feedback and interaction with 
physicians, nurses, pharmacists, and others has been vital to the ongoing success and value of these systems.

In this panel, leaders from these different efforts explore the practical challenges, design and deployment 
innovations, implementation strategies, rollout experience, real achievements and ongoing issues from these efforts. 
The panelists will discuss lessons learned in the process that can make it possible to be more prepared for, and more responsive to, crises yet to come.

**Participants**

**Jonathan Teich, MD, PhD, FACMI** will moderate the panel and discuss functional design considerations, 
optimization of speed and usability for specialized care settings, and clinical decision support (CDS) opportunities, 
based on practical clinical considerations as well as informatics principles. Dr. Teich has over 20 years of experience 
in EMR design, CDS and rapid access to medical knowledge, working in academic, provider, industry and 
government settings; he is a practicing emergency physician and has studied and written on disaster and crisis 
informatics since 2001.

**Hamish Fraser, MBChB, MSC, FACMI** will discuss the wider information needs of disaster management, 
resource and supply chain management, and working with local and national leaders. He will also discuss evaluation 
and research to ensure effective and usable systems that work even in crisis situations, and how all the organizations 
doing crisis response work can collaborate further on these eHealth systems and implementation strategies. Dr. 
Fraser has worked for over a decade to bring medical informatics tools and expertise from developed countries to 
some of the most challenging environments in the developing world, and is currently working to develop a 
framework and strategies for evaluation of these tools worldwide.

**Eric Peraklis, PhD** will discuss needs assessment, systems selection and similarities and differences of the various 
platforms and available options to provide needed capability, with a particular focus on contact tracing, research and 
infrastructure at the MSF treatment centres. Dr. Peraklis is an informatics and R&D leader with a background 
ranging from pharmaceutical discovery to cybersecurity to bioinformatics. He serves as executive director of the 
Center for Biomedical Informatics at Harvard Medical School, and has had leadership positions at the FDA and at 
Johnson & Johnson.

**Shefali Oza, MSc** will discuss the requirements for efficiently implementing technology-based solutions during a 
disaster. Besides the software and hardware, successful implementation requires planning, management, training, 
continuous feedback, and a staged rollout, all of which present unique challenges during complex humanitarian 
emergencies; she will discuss these issues in the context of this Ebola EMR deployment. Ms. Oza is an 
epidemiologist at the London School of Hygiene and Tropical Medicine and has worked on global health projects in 
five continents.
Darius Jazayeri, MEng, will discuss the software, system infrastructure, development and maintenance considerations of the functional modules as well as the OpenMRS platform that underlies both of these projects. He is a Principal Architect with the Global Health team at ThoughtWorks, Inc, and a Software Developer with the Medical Informatics team at Partners In Health. He has built eHealth systems for developing countries for 15 years, including clinical data capture, drug order entry, reporting, analytics, and more. He is a lead developer of the OpenMRS open source medical record platform, where he is responsible for the architecture of its reference application, and helps coordinate contributions from developers worldwide.

Importance to the Audience

This panel is clearly urgent and timely; Ebola was a ferocious killer in 2014 and while the numbers are reduced as of this writing, the epidemic is certainly not over, and could flare up again. Moreover, we can be sadly confident that more epidemics, disasters and health crises will occur over time. Information technology is clearly a vital part of the effort to halt these rampages, and it just as clearly requires new ways of thinking in design and implementation, particularly when the crisis requires hazardous or difficult working conditions.

Because there will be more crises, we should pay attention now to shared lessons, common core infrastructure, necessary modular functions and readiness planning, so that the time from the start of the next crisis to the implementation of usable and valuable information tools can be much shorter. The panel will bring new ideas and show the way to new shared efforts that can be started now. In addition, the audience will gain additional motivation and strategies for including research and epidemiology in any project so that care can improve.

Even beyond any health crisis, the innovations developed in Africa may have implications for informatics and system design everywhere. In the process of developing easier systems for difficult environments, the teams have found techniques that lead to faster and easier systems in general. Lessons learned from the Ebola work can be applied now to better daily-use EMRs in “normal” environments.

The target audience includes system designers, implementation leaders, CMIOs and CNIOs, global health leaders, health services researchers, students and anyone who is interested in what has been done to combat this lethal epidemic.

Discussion Questions

1. How do you work with the local governments and health bodies to ensure success?
2. How can you ensure robust maintenance, support and necessary system enhancements?
3. How can we collaborate to harmonize these and other efforts, to share strategies and toolkits, and in general to make these innovations more widely and available?
4. What types of research are most valuable while simultaneously practical and achievable in these situations?
5. What other new devices, technologies, designs or strategies may be useful to address the challenges?
6. What can we do to have a quicker, better response to the next crisis?

Panel Organizer’s Statement

All participants have contributed to the proposal and agree to take part on the panel.

References

State of the Art of Clinical Narrative Report De-Identification and Its Future

Peter Szolovits, PhD, 1 John Aberdeen, MA, MS, 2 Stephane Meystre, MD, PhD, 3,4 Mehmet Kayaalp, MD, PhD 5

1Massachusetts Institute of Technology, Cambridge, MA; 2MITRE Corporation, Bedford, MA; 3University of Utah, Salt Lake City, UT; 4VA Health System, Salt Lake City, UT; 5National Library of Medicine, Bethesda, MD

Abstract

While automatic de-identification systems exist, release of de-identified data usually requires significant multi-round expensive effort for validation. To overcome this barrier, we need a consensus on the parameters of successful automatic de-identification. Although we can establish such parameters relative to error rates of human annotators, it is ultimately a policy question whose answer needs to be vetted by the public.

When personal identifiers are substituted with surrogates or pseudonyms, it could be very difficult to spot the residual identifiers missed by the de-identifier, but in absolute terms, it is difficult to ensure that de-identified clinical text contains no references that might indirectly identify the patient; hence, de-identified clinical text is usually shared through a data use agreement. When such an agreement is in place, dates and some address parts can be left identified in a limited data set. If we reframe the de-identification problem by focusing on the pertinent identifiers, we may smooth the path to data sharing.

Installing and running a clinical de-identification system may require substantial expertise, which small institutions and clinical scientists may lack. Although we would like to develop capable systems with sophisticated functionalities, we also should strive for simplicity for routine de-identification tasks.

Audience

- Clinical scientists
- Clinical practitioners
- Clinical informaticians
- Clinical application developers
- Policy makers in health institutes

Questions

1. How can we reach out to policy makers to establish parameters of successful automatic de-identification?
2. How should we approach the public, explain the importance of the availability of de-identified clinical narrative reports for the advancement of clinical sciences, and win the public trust?
3. How can we ensure the integrity of information, when we pseudonymize de-identified text? What should be the safety margins?
4. Can de-identification be provided as an online service? If so, how can we solve the trust issue and overcome institutional barriers?
5. How can we reach out to clinical institutions and convince clinicians not to use personal identifiers in their clinical reports?
Why the topic of this panel is timely and/or controversial

Clinical text de-identification (CTD) is no longer a dream. It is an increasingly important practical tool for making clinical text data available to a wider clinical research community. This panel will discuss a number of hot topics surrounding CTD. There is no default answer to any of those questions. Many of those issues are quite controversial. For example, CTD as an online service is an attractive but untested idea and it could be the only practical solution for many small institutions with limited expertise in informatics.

Information about panelists and their topics of discussion

Peter Szolovits is Professor of Computer Science and Engineering in the MIT Department of Electrical Engineering and Computer Science (EECS) and an Associate faculty member in the MIT Institute of Medical Engineering and Science (IMES) and its Harvard/MIT Health Sciences and Technology (HST) program. He is also head of the Clinical Decision-Making Group within the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). His research centers on the application of AI methods to problems of medical decision making, natural language processing to extract meaningful data from clinical narratives to support translational medicine, and the design of information systems for health care institutions and patients. He has been a founder of and consultant for several companies that apply AI to problems of commercial interest. He received his bachelor's degree in physics and his PhD in information science, both from Caltech. Prof. Szolovits was elected to the Institute of Medicine (now the National Academy of Medicine) and is a Fellow of the American Association for Artificial Intelligence, the American College of Medical Informatics and the American Institute for Medical and Biological Engineering. He recently served as a member of the National Research Council's Computer Science and Telecommunications Board and is a member of the National Library of Medicine’s Biomedical Library and Informatics Review Committee. He is the 2013 recipient of the Morris F. Collen Award of Excellence from the American College of Medical Informatics.

Making clinical data available for research is an absolute necessity for improving our understanding of disease and raising the quality of decisions in health care. However, that benefit is offset by risks to confidentiality of patient data. Stripping or replacing explicit identifiers such as names, medical record numbers, addresses and the other HIPAA-specified data elements is straightforward, but de-identifying narrative reports accurately would require automated procedures almost to understand the content of those reports. Automated de-identification tools have taken several main approaches to protect confidentiality without sacrificing too much of the value of the narrative data. Rule-and-dictionary based methods can identify both specific elements that are to be elided (e.g., names) and elements that are to be retained (e.g., medications), but can err on out-of-vocabulary elements. Machine learning methods require substantial training on annotated data and need retraining when applied to new data sets where stylistic and linguistic conventions may differ. Synthesis techniques model the distributions and relationships among data elements and synthesize new data sets with similar characteristics, which assure that no actual data can leak; but they support well only those analyses that rely on the same retained characteristics. Differential privacy, thus far mostly applied to coded rather than narrative data, may provide another approach, but tends to destroy much data in order to maintain its confidentiality guarantees. The choice of the “best” techniques to apply are probably application dependent.

John Aberdeen is the project leader for the open-source MITRE Identification Scrubber Toolkit (MIST), which provides an environment to support rapid tailoring of automated de-identification to different record types, using automatically learned classifiers. He will discuss recent research about the replacements used when de-identifying clinical notes, and how careful surrogate generation can render de-identified notes more resistant to re-identification attacks.

In contrast to rule-based systems, de-identification via trained classifiers provides a clean separation between a clinician’s domain knowledge (applied when generating training data for the classifiers), and the software that
applies that knowledge. All de-identification processes, when manual or automated, leave behind some (often small) number of un-redacted PHI elements (residual PHI), with an associated risk of exposure. Rather than attempting in vain to make tiny incremental improvements to eliminate residual PHI, we can accept that de-identification is imperfect, and carefully choose PHI replacements that blend smoothly in with the surrounding text. If we do this well, the residual PHI become extremely difficult to detect (hiding in plain sight), and cannot be used to launch a re-identification attack.

**Stéphane Meystre** is faculty member of the Department of Biomedical Informatics at the University of Utah with research activities focused on easing access to clinical data for research and clinical care purposes, using techniques such as Natural Language Processing for information extraction and automated de-identification, and automating ontologies development and management.

The U.S. Department of Veteran’s Affairs (VA) is funding an informatics initiative called the Consortium for Healthcare Informatics Research (CHIR), focused on utilizing both structured and unstructured data previously unavailable for research and operational purposes. Evaluating existing de-identification methods and building and evaluating new methods and tools is one of the cornerstones of this initiative, and will be presented during this panel. Realized efforts shall fulfill the ethical and legal obligations of patient privacy and confidentiality.

Dr. Meystre is leading the CHIR de-identification project, building a best-of-breed de-identification system for VA clinical documents, evaluating its interaction with subsequent text processing, and eventually how anonymous de-identified documents are. He will first introduce the audience to the best-of-breed system developed for VA clinical text, and share issues related with text de-identification. He will then present generalizability and adaptation to VA text difficulties.

**Mehmet Kayaalp** is the organizer of the panel. He graduated from University of Istanbul with an M.D. degree, from Southern Methodist University with an M.S. degree in Computer Science and from University of Pittsburgh with a Ph.D. degree in Intelligent Systems. Since 2003, he is in public service working at the National Library of Medicine as staff scientist. He currently leads there the NLM Scrubber project. NLM Scrubber is a clinical text de-identification system, freely available at [https://scrubber.nlm.nih.gov](https://scrubber.nlm.nih.gov).

Mehmet will discuss various aspects of the following questions:

- Is there a better way to generate de-identified data? Downloading and running stand-alone clinical text de-identification applications is the current standard for sharing. What other options are there? How feasible is it to distribute large-scale de-identified clinical text data? Would small clinical research institutes be open to de-identify their clinical reports through an online service?
- Can we go beyond the token-based precision and recall metrics to measure the actual performance of a de-identification system? What are the problems of the current evaluation approaches and how can we remedy those?
- Should we redefine our goals of success, re-prioritize the de-identification tasks at hand, and focus on the personal identifiers that need to be de-identified in limited data sets?

All participants have agreed to take part on the panel.
Informatics Research and Innovation in a Commercial Electronic Health Record:
The Experience of Three Organizations Using Epic
Organizer: Adam Wright, PhD

Panelists:
Adam Wright, PhD a,b,c
David W. Bates, MD, MS a,b,c
Eric S. Kirkendall, MD, MBI d,e
David A. Dorr, MD, MS f
Peter DeVault, MS g

Panelist Affiliations:
a Partners Healthcare System, Boston, MA
b Brigham & Women’s Hospital, Boston, MA
c Harvard Medical School, Boston, MA
d Cincinnati Children’s Hospital Medical Center, Cincinnati, OH
e University of Cincinnati, Cincinnati, OH
f Oregon Health & Science University, Portland, OR
g Epic Systems, Verona, WI

Abstract: (150-200 words)
Historically, many of the key innovations in informatics came from organizations with self-developed clinical information systems. Having a self-developed system gives organizations considerable flexibility to tailor and extend their software and explore new and potentially innovative approaches. However, maintaining a self-developed system is expensive, poses risks to generalizability and makes scaling challenging. Recognizing these issues, most academic medical centers have moved, or are in the process of moving, to commercial electronic health records (EHRs).

In this panel, we present the experience of three organizations which are actively implementing innovative approaches to improve care: Partners HealthCare, Cincinnati Children’s Hospital Medical Center and Oregon Health and Science University. All three organizations, at one time, had pioneering self-developed systems but switched to Epic, one of the most popular EHRs used worldwide.

The presenters will detail their organizations’ experience innovating on the Epic platform through the lens of several case studies, introduce the range of technical approaches for developing and integrating innovative tools in Epic, discuss Epic’s approaches to working with innovators and cover issues in organization governance of innovation activities.

Though the focus of the panel is the experience with Epic, the lessons learned apply to all commercial EHRs.
Description:
The overall outline of the talk is given in the abstract. The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>Wright</td>
<td>Introduction, overview of innovation in self-developed and commercial systems</td>
</tr>
<tr>
<td>15 min</td>
<td>Wright</td>
<td>Overview of approaches to innovation in Epic and the Partners experience</td>
</tr>
<tr>
<td>10 min</td>
<td>Kirkendall</td>
<td>Cincinnati Children's Hospital Medical Center experience</td>
</tr>
<tr>
<td>10 min</td>
<td>Dorr</td>
<td>OHSU site experience</td>
</tr>
<tr>
<td>10 min</td>
<td>DeVault</td>
<td>Epic’s approach to training and working with innovators</td>
</tr>
<tr>
<td>10 min</td>
<td>Bates</td>
<td>Lessons learned and innovation governance</td>
</tr>
<tr>
<td>30 min</td>
<td>Panel</td>
<td>Discussion and questions from audience</td>
</tr>
</tbody>
</table>

The specific topics covered by each presenter will be:

1. Wright (introduction): Introduce the topic and panelists, including a brief description of what each panelist will be talking about and their affiliations. Provide an overview of the history of informatics innovation in self-developed systems and the challenges and opportunities of innovation in commercial systems.
2. Wright (overview and Partners experience): Give an overview of the modes of integration for innovations in Epic, including data extracts, web services, building clinical content, embedding external applications and Open.epic services, including FHIR. Present the Partners experience innovating in Epic.
3. Kirkendall: Give an overview of Cincinnati Children's Hospital Medical Center’s experience innovating in and extending Epic, including challenges and lessons learned.
4. Dorr: Give an overview of OHSU’s experience innovating in and extending Epic, including challenges and lessons learned. Dr. Dorr’s presentation will be coordinated with Dr. Kirkendall’s presentation to ensure they do not overlap.
5. DeVault: Give an overview of Epic’s resources for innovators, including training, certification and support, as well as Epic’s approach to innovation.
6. Bates: Summarize and synthesize the experience of the panelists, the challenges and lessons they encountered along the way, concluding with a vision for approaches to innovation governance that organizations can adopt.
7. Panel: After the presentations, the panel will briefly discuss additional issues and then open for questions from and discussion with the audience. Audience members will be invited to address questions to specific panelists, or to the panel as a whole.

Importance:
As many organizations shift from self-developed EHRs to commercial systems, and many academic institutions have switched to Epic in particular, questions have been raised in a variety of venues about how informatics research and innovation will be able to proceed. Many participants have expressed concern that commercial systems may be inflexible, stifling innovation. However, Epic and other vendors are increasingly supporting a variety of approaches that enable innovation, including additional approaches for customizing clinical decision support and open Application Programming Interfaces (APIs).
In this panel, we outline the experience of three organizations that have successfully innovated in Epic, and also describe the range of approaches for innovation. As organizations continue migrating to commercial systems, we believe it will be essential for the informatics community to engage with commercial systems, and develop best practices and approaches for innovating on commercial platforms. We believe that this panel will be critical to advancing this discussion and providing a roadmap for organizations and individuals looking to innovate on commercial platforms.

Discussion Questions:
1. What tools and processes can EHR vendors provide their customers to enable innovation?
2. What are best practices for encouraging and fostering innovation in clinical organizations?
3. How can organizations develop and employ effective governance approaches for informatics innovation?
4. What are the similarities and differences in informatics research and innovation in self-developed and commercial systems?
5. How can innovations developed at one site be effectively disseminated to other organizations using the same or different software?
6. What were the challenges and lessons learned encountered by the sites as they developed innovation approaches?
7. What are future directions and research and innovation foci for sites using commercial systems?
8. How do you engage and communicate with vendors in regards to innovation?
9. What are the risks and benefits of innovating with vendor software?

Participation statement:
All proposed panelists are aware of this panel submission, and have agreed to participate in the panel if the proposal is accepted.
Challenges, Successes, and Future Directions of Consumer Health IT Evaluation

Uba Backonja, PhD, RN\(^1\) (organizer); Rupa Sheth Valdez, PhD\(^2\) (moderator); William Riley PhD\(^3\), Katie Sick, PhD\(^4\); Teresa Zayas-Cabán, PhD\(^5\); Kenneth Goodman, PhD\(^6\)

\(^1\)School of Medicine, University of Washington, Seattle, WA; \(^2\)Department of Public Health Sciences, University of Virginia, Charlottesville, VA; \(^3\)Science of Research and Technology Branch, National Cancer Institute, Bethesda, MD; \(^4\)Informatics Division, Indiana University, Bloomington, IN; \(^5\)Health IT Portfolio, Agency for Healthcare Research and Quality, Rockville, MD; \(^6\)Bioethics Program, University of Miami, Miami, FL

Abstract

The accelerated growth and diversity of consumer health IT interventions creates an urgent need to develop evaluation criteria and methods to assess their effectiveness and efficacy. This panel will discuss challenges, successes, and future directions of evaluating consumer health IT interventions. Panelists will discuss their evaluation experiences: Dr. Riley will discuss accounting for the rapid pace of technological change; Dr. Siek will discuss accounting for patients’ workflow and daily living contexts; Dr. Zayas-Cabán will discuss accounting for intervention efficacy and safety; Dr. Goodman will discuss accounting for ethical considerations. The session will then serve as a forum for the panelists and audience to discuss topics related to consumer health IT evaluation, such as how to: modify existing knowledge and evaluation frameworks for consumer health IT evaluation; assess the safety and cost/benefit of the interventions; and balance evaluation rigor and richness with the need to efficiently evaluate rapidly changing technologies. This panel is timely given the rapid uptake of consumer health IT technologies, which necessitates the development of new and innovative methods to evaluate these technologies and interventions.

Intended Audience

Individuals who have experience or interest in consumer health IT research and/or intervention evaluation.

Aim of the Discussion

To discuss the challenges of and strategies for evaluating consumer health IT interventions.

Overall Panel Topic

Consumer health technologies and interventions are becoming ubiquitous. However, existing evaluation criteria and methods to assess whether consumer health IT interventions are safe and effective are still in their infancy. For example, only recently (February 2015) did the Food and Drug Administration release revised guidance on its oversight of mobile health (mHealth) applications. Therefore, the purpose of this panel is for panelists to discuss their experiences in consumer health IT evaluation, and to converse with the audience about how to address consumer health IT evaluation challenges.

Contribution of Each Panelist

At the start of the session, each panelist will provide a short presentation about his or her work in consumer health IT evaluation.

William Riley, PhD

Dr. Riley will focus on the need for consumer health IT evaluations to account for the rapid pace of technological change. Consumer technologies change more rapidly than technologies that need to be selected and implemented by organizations, and a similar relatively rapid pace can be expected in consumer health IT. Consequently, evaluation of consumer health IT before, during, and after its adoption by consumers needs to be conducted in a highly timely manner while maintaining a high level of rigor. The challenges of conducting evaluations that keep up with rapid technological change include the trade-off of evaluation rigor and evaluation speed, changes in evaluation criteria and relevant measures due to new information and changes in the industry, changes in the technologies under evaluation themselves, and the likelihood of obsolescence of delayed evaluation findings. Strategies that will be
discussed include early and frequent integration of evaluation with design using empirically-based design and single case study designs, and integration of evaluation within interventions and dissemination activities, with automated randomized controlled trials and quasi-experimental study designs.

Katie Siek, PhD
Dr. Siek will focus on the need for consumer health IT evaluations to account for patients’ workflow and the daily context of the home and community. Although consumer health IT has the potential to benefit patients and other lay users by providing and facilitating the generation of personalized health information, these technologies also have a high abandonment rate. Consequently, we must understand how dimensions of people's daily lives (the technologies they use, their physical and social environments, and their routines) interact with specific consumer health IT interventions to facilitate or hinder adoption, use, and appropriate outcomes. The challenges of accounting for patients’ workflow and daily contexts include gaining long-term access to patient populations and the environments in which consumer health IT solutions are intended for use and an in-depth understanding of patients’ use experiences. To this end, we propose borrowing heavily from Community Based Participatory Research methods and combining them with well-established user centered design and participatory design techniques.

Teresa Zayas-Cabán, PhD
Dr. Zayas-Cabán will focus on the need for consumer health IT evaluations to account for intervention efficacy and safety. Evaluating efficacy of consumer health IT is critical prior to implementation to ensure it works as intended and has the desired effect. The design and implementation of health IT can impact how the system is used and lead to errors, and evaluating the safety of systems patients are expected to utilize “on their own” is critical to avoid causing harm. Challenges to evaluating efficacy and safety of consumer health IT include intervention fidelity, difficulties detecting intervention impact, and distinguishing it from other effects (e.g., patient behavior, concurrently implemented quality improvement programs). When conducting field-based studies there may also be recruitment, enrollment, and follow-up challenges. Strategies to overcome these challenges include a strong recruitment plan with referrals from local organizations (e.g., health sites), a robust design process that contains testing prior to implementation, and leadership buy-in when consumer health IT is being implemented through a clinical or other organization.

Kenneth Goodman, PhD
Dr. Goodman will focus on the need for consumer health IT evaluations to account for ethical considerations. Evaluation itself is an ethical obligation, which must be attended to both before and after implementation. Unfortunately, this ethical obligation to engage in evaluation is often overlooked by the medical informatics community, including when interventions are oriented toward consumers. The challenges of consumer health IT evaluation themselves include ethical issues, including that of consent and privacy. To address these challenges, the consumer health informatics community should borrow approaches to trusted governance, which can have the effect of fostering trust in the communities of patients and other lay users whose health IT use we seek to evaluate. Community engagement then emerges as a strong, useful and symbiotic tool.

Panel Topics
After the introductory presentations, the panel and audience will engage in a discussion of various topics, facilitated by the moderator. Topics relevant to this panel are:

• What are the major barriers to conducting consumer health IT research and evaluation?

• What are common privacy and security concerns that arise during the community-based evaluation and research of consumer facing consumer health IT?

• How do we promote research in community-based health IT interventions, especially those developed by the private sector, not academia, and that do not require FDA approval for commercialization?

• Is a balance possible between (1) streamlining the evaluation process, which may take a long time especially in an academic setting, to keep up with the fast pace of technological change and innovation, and (2) ensuring the evaluation is thorough enough to ensure the safety of patients and their personal health information, and the efficacy of the intervention?
• What are commonalities across different strategies to evaluate consumer health IT?
• How can we engage patients and informal caregivers in evaluating mHealth applications?
• How does evaluation of consumer health IT differ from evaluation of clinical health IT?
• What evaluation techniques for consumer health IT can be borrowed from other domains?
• What challenges exist to developing generalizable knowledge for consumer health IT evaluation?

After the discussion of topics regarding evaluation of consumer health IT, the moderator will lead the panelists in a debate to resolve the following question: Is it possible to identify a unified framework for research and evaluating consumer health IT, or are unique research and evaluation methods needed for different types of consumer health technologies?

Goals of the Panel Discussion
(1) Foster an interdisciplinary dialogue related to consumer health IT evaluation; (2) Provide AMIA symposium attendees with an understanding of multiple consumer health IT evaluation frameworks that they may leverage in their research and practice; (3) Provide AMIA symposium attendees with an understanding of current challenges in consumer health IT evaluation and a call to action regarding future efforts required.

Panel Timeliness
This panel is timely because of the rapid proliferation of consumer health IT. Evaluating the effectiveness of new waves of patient-facing technologies will require merging existing evaluation frameworks with new measures that are specific to novel contexts. Because of its interactive nature, this panel will serve as a forum in which members and panelists can collaboratively advance dialogue related to consumer health IT evaluation. Specifically, audience members will have the opportunity to learn from the panelists and one another about methodological and field-based challenges and innovations that others have encountered and developed. This sharing of experiences will then serve as a foundation for a discussion of future research directions needed to advance the science of consumer health IT evaluation.

All panelists have agreed to take part on this panel.
Patient Portals: Best Practices and New Directions for Development and Investigation

Patricia Dykes, RN, PhD, MA 1, Sarah Collins RN, PhD 1,2, Anuj Dalal MD 1, Cindy Dwyer, RN, BSN 3, S. Ryan Greysen, MD, MHS, MA 4

1 Brigham and Women’s Hospital and Harvard University, Boston, MA
2 Partners HealthCare, Boston, MA
3 Johns Hopkins Hospital, Baltimore, MD
4 University of California San Francisco School of Medicine, San Francisco, CA

Abstract

The meaningful use legislation requirements for patient engagement have led to a greater use of personal health records or patient portals in outpatient settings. However, there are limited examples in the literature that describe strategies and e-health tools to provide patients with access to their health information in hospital or inpatient settings. This interactive panel will explore the findings from a recent study that identified inpatient portal best practices across four “early adopter” healthcare organizations in the United States. Session participants will be involved in refining and identifying new directions for the development and investigation of patient portals.

General Description

Many consumers are using technology to manage health and wellness, and there is a growing recognition of the need for patient engagement in healthcare. The Health Information Technology for Economic and Clinical Health (HITECH) Act meaningful use program (2009) requires that providers engage patients in their healthcare through the use of technology. A key goal of meaningful use is to make consumers full partners in their care by providing e-health tools that increase access to health information, support activation (e.g., active involvement in their treatment plan), and that help consumers to gain control over their health and wellbeing (1, 2). To accomplish this, providers and healthcare organizations have promoted the use of personal health records or patient portals, typically in ambulatory settings (3, 4). There are limited examples in the literature that describe strategies and e-health tools to provide patients with access to their health information in acute care settings. Evaluation of current state of patient portals in acute care settings is needed to inform the development of best practices, to raise awareness of issues and challenges encountered by early adopters, to encourage standardization of features provided to patients, to promote further investigation, and to spur innovation (1).

The LIBRETTO Consortium, funded by the Gordon and Betty Moore Foundation, is exploring best practices for acute care patient portals across four “early adopter” healthcare organizations in the United States. The key goals of the project are to 1) identify key features of an acute care patient portal that are common across sites, 2) recognize common operational and implementation challenges (e.g., workflow, vendors, mobile device security/infection control, etc.), 3) define best practices for acute care patient portal design, configuration, and use (e.g., best practices for patient-provider messaging, delivering educational content, etc.), and 4) propose new directions for investigation and innovation specific to the acute care settings (intensive care and non-intensive care), focusing primarily on personalizing and enhancing the patient/family caregiver experience.

This interactive panel will engage participants in a range of key topics and share lessons learned from the LIBRETTO project. Participants will be encouraged to identify and share additional lessons learned from their institution.

Discussion Topics

For each of the discussion topics below, the panel will present a brief update on LIBRETTO findings and engage the audience in a discussion related to additional experiences and best practices.
What are the common and unique features at each site that serve to engage patients and their care partners while hospitalized in the acute care/ICU setting?

What are key operational and implementation challenges that were encountered? From the patient/caregiver perspective? From the provider perspective?

How does each site plan to enhance the experience of patients and family caregivers using the portal? What is the content that providers want to know about their patients? What information do patients wish to share with providers?

What are the privacy/legal ramifications of providing patients/health care proxies the ability to share aspects of the patient’s plan of care (e.g., schedule) with family members?

How do we deliver educational content most effectively using multimedia modalities?

How do we leverage patients’ social networks and existing social media profiles to set-up their hospital profile to share with providers?

How do we incorporate video conferencing technology to facilitate goals of care discussions with off-site family caregivers and providers?

How do we leverage the full potential of mobile devices? What are the advantages and disadvantages of a hospital-issued device versus bring your own device (BYOD) strategy?

Moderator: Patricia Dykes, RN, PhD, MA is Senior Nurse Scientist and Program Director for Research in the Center for Patient Safety Research and Practice and the Center for Nursing Excellence at Brigham and Women’s Hospital (BWH) and Assistant Professor at Harvard Medical School. Her research interests are in quality and safety of care and adverse event prevention, especially as it relates to hospitalized inpatients. While funded by the Robert Wood Johnson Foundation, Dykes and team developed a fall prevention toolkit that significantly reduced falls in hospitals. She has expanded this research to explore the use of technology to provide the core set of information needed by patients and care team members to engage in safe patient care. Dr. Dykes is a fellow of the American Academy of Nursing and the American College of Medical Informatics. She is the author of two books and over 50 peer-reviewed publications, including the 2013 article in Journal of Gerontological Nursing, Building and testing a patient-centric electronic bedside communication center, which describes the use of a participatory design process to develop and test an inpatient portal prototype to improve access to health information for hospitalized adults and their family caregivers. The prototype was recently enhanced to include additional tools to support patient engagement in their plan of care and is in use by patients and family caregivers on iPads on the medical intensive care and oncology units at BWH.

Panel Members

Sarah Collins RN, PhD is a Clinical Informatician in Clinical Informatics Partners eCare at Partners Healthcare Systems and an Instructor in Medicine at Harvard Medical School and Brigham and Women’s Hospital (BWH) Division of Internal Medicine and Primary Care. Dr. Collins is an experienced critical care nurse. Her research, as well as her applied clinical informatics work, is focused on modeling, developing, and evaluating standards-based, patient-centered collaborative informatics tools to further patient safety and coordinated patient-centered care. Dr. Collins was first author of a 2011 JAMIA publication titled: Policies for Patient Access to Clinical Data via PHRs: Current State and Recommendations. She is leading the MySafeCare research project at BWH. MySafeCare is an application for patients and care partners to electronically submit safety concerns while in the hospital for real-time review via clinical dashboards. This AHRQ funded work includes integration of MySafeCare with the Acute Care Patient Portal at BWH. Her prior patient safety acute and critical research has found significant associations between nursing documentation and patient outcomes of cardiac arrest and mortality development. Dr. Collins’ foundational research aim at increasing patient-centered and collaborative care in the inpatient setting includes development of a conceptual model of Interdisciplinary Information Exchange in the Intensive Care Unit, an Interdisciplinary Handoff Information Coding Framework extending the Continuity of Care Document (CCD) standard for the inpatient setting, and content and functional specifications for a standards-based multidisciplinary rounding tool in acute and critical care.
Anuj Dalal, MD is an Associate Physician and a member of the BWH Hospital Medicine Unit within the Division of General Medicine and Primary Care at Brigham and Women’s Hospital (BWH). He is an Assistant Professor at Harvard Medical School and Fellow of Hospital Medicine. He completed the Program in Clinical Effectiveness at Harvard School of Public Health, and holds a Graduate Degree in Medical Informatics from Oregon Health & Science University. He has expertise in designing, implementing, and evaluating innovative HIT strategies to improve the delivery of care in the inpatient and transitions setting. Specifically, he has worked on initiatives to improve communication of clinically significant test results, optimize care team communication, and facilitate patient engagement with an overall goal of improving patient safety, quality, and healthcare costs. He is co-leading the PROSPECT (Promoting Respect and Ongoing Safety through Patient-centeredness, Engagement, Communication, and Technology) initiative at BWH with Dr. Dykes. The primary goal of this initiative is to engage patients and care partners in the acute care setting implementation of a novel, web-based patient-centered toolkit (PCTK) in the acute care setting. He is leading the effort at implementing a patient-centered microblog that serves as the web-based and mobile messaging platform for the PCTK.

Cindy Dwyer, RN, BSN is a Nurse Clinician III in the Surgical Intensive Care Unit at The Johns Hopkins Hospital. She has worked recently with the Armstrong Institute for Patient Quality and Safety and the Johns Hopkins University Applied Physics Lab in the implementation of the Emerge project. Cindy has combined her over 20 years of critical care bedside nursing with experts in Patient and Family Centered Care to assist in the development and evaluation of a Patient and Family Portal as part of project Emerge. She has completed the Armstrong Institute Patient Safety Certificate Program and is a member of the American Society of Professionals in Patient Safety. Cindy is currently exploring innovative approaches to enhancing the patient and family experience in the intensive care setting.

S. Ryan Greysen, MD, MHS, MA is Assistant Clinical Professor in the UCSF School of Medicine. His research focus is transitions of care for hospitalized older adults and interventions to improve post-discharge continuity of care including novel uses of patient-centered technology such as mobile devices and social media. Dr. Greysen’s mixed-methods evaluation studies have leveraged advanced quantitative and qualitative techniques to inform health policy at the intersection of mobile health, hospital medicine, and geriatrics. His 2014 publication in the Journal of Hospital Medicine shows that hospitalized patients can be more engaged in their care by using tablets to access their patient portal at the bedside and that older or less tech-savvy patients benefit just as much as others if given focused coaching. His ongoing work expands this bedside coaching with portals and integrates mobility sensors to engage older adults/caregivers in goal-setting to avoid complications such as functional decline during hospitalization.

Acknowledgements: We acknowledge The LIBRETTO Consortium, funded by the Gordon and Betty Moore Foundation, for supporting this work. Additionally, we thank Dr. David Vawdrey for participating in our survey of Acute Care Patient Portals Best Practices and his continued work in this space and collaboration in this research.

References

The Informatics Sculptor & the Clinical Annotator: Effective Annotation Strategies

Panelists

Nancy Gentry, RN, MSN1,6, Elizabeth Hanchrow, RN, MSN1,6, Glenn T. Gobbel, DVM, PhD, MS1,6, Brett R South, MS2,3, and Steven M. Bradley, MD, MPH4,5

Organizer & Moderator

Ruth Reeves, PhD1,6

1Tennessee Valley Healthcare System VA, Nashville, TN; 2 University of Utah, Salt Lake City, UT; 3Salt Lake City VA Medical Center, Salt Lake City, UT; 4Eastern Colorado Health Care System VA, Denver, CO; 5University of Colorado School of Medicine Denver, CO; 6Vanderbilt University, Nashville, TN

Abstract Annotation of free text is an operation requiring human labor, complex cognition and topic-specific knowledge, requiring multiple skills sets in order to be effectively performed. This panel organized around the shared perspective that cross-education of research team members in the disciplines involved is a necessary and desirable function of informatics research. Curating clinical data thru analytic processes including annotation depends on multiple levels of data transform, each requiring expertise from the participating roles involved in the project. This panel discussion includes the following roles: 1) annotator, 2) data quality organizer, 3) NLP developer, 4) clinical expert study designer. Panel members representing one of the roles in integrating annotation data into the research project will discuss tasks and responsibilities of their own role within the process, converging on how their own professional perspective impacts and is impacted by the other roles in the project. The barriers to operationalizing annotation goals will be significantly lowered for attendees of this panel session. Those with an interest in initiating an NLP system for their research or in obtaining quality annotation for other purposes will gain an understanding of ways to best make use of each team member’s particular expertise to obtain quality annotations.

Introduction Annotation of free text is an operation requiring human labor, complex cognition and topic-specific knowledge. In the context of clinical informatics research, it is the process of identifying discrete and often highly variant pieces of textual data (typically within narrative clinical documents) and classifying them for the purposes of establishing normalized equivalence classes for machine-readable processes. More complex annotation tasks may provide signals that aid natural language processing (NLP) systems in performing automated inferencing tasks such as assigning negation, probability scales, chronological ordering or establishing relationships between identified concepts and entities. Traditional View of Annotation in Clinical Research: Medical research projects requiring substantial data extraction from clinical notes often expend substantial resources and time, contracting human reviewers to annotate medical corpora with the relevant clinical concepts and attendant attributes. Historically, such projects recruited physicians. However, this approach has proven expensive, and it is difficult to maintain annotator engagement over the life-time of a project due to competing educational or work-place demands. Alternatively, researchers have attempted acquiring development data by employing non-medically trained annotators aided by clinical subject matter experts, resulting in requiring numerous training cycles and only modest inter-annotator agreement rates. Emergent View of Annotation in Clinical Research: The traditional perspective on annotation labor views it within the pyramid model, with the principal investigator as the author of the study design at the top, the development team as middle managers, and annotators as data entry labor at the bottom. A more innovative approach is a flat-structured interdisciplinary model which organizes around the shared perspective that cross-education of research team members in the disciplines involved (e.g., adverse event documentation practices, NLP, nephrology) is a necessary and desirable function of informatics research.

Discussion Aims The traditional approach to annotation in NLP research is costly, inefficient, and suboptimal. We propose a new approach that fosters interdisciplinary contribution and efficiency. The goal of this panel is to move the audience to see the advantages of annotating within a flat-structured interdisciplinary research setting and to illustrate how to obtain highly reliable and quality annotation data as a function of the overall research project aims. Panelists will comment on and invite discussion on the points of tension involved in annotation procedures in real-life informatics research projects they have been involved in, with a focus on how best to marshal these for producing high quality annotations. The panel will be comprised of four researchers and a moderator who have each had substantial experience with developing or making use of annotation projects in an informatics research operation. Each panel member will address one of the necessary roles in integrating annotation data into the research project: 1) the annotator, 2) the data quality organizer, 3) the NLP developer, and 4) the clinical expert study designer. Each will discuss the tasks and responsibilities of their own role within the process, and converge on how these tasks and his or her own professional perspective impacts and is impacted by the other roles in the project.

252
The presentations will describe how each role trains the other roles, i.e. learning is not one-directional from research PI or clinical expert to annotator.

**Anticipated Audience for Topic of Controversy** Audience participants, either new to or experienced in NLP research, will learn new ways to reduce study risks and costs and associated angst for more productive methods in operationalizing annotation. Investigators with an interest in obtaining quality annotation will gain an understanding of ways to best make use of each team member’s particular expertise. Annotation is often a first step for informatics research projects that include extracting data from unstructured text. In the experience of the panelists and moderator, informatics projects that use annotation require an interdisciplinary approach; leveraging the cumulative knowledge of the annotators and the consumers of the annotation products generates critical contributions to the informatics science of the project. This view veers from the prevailing one of annotation as a necessary burden, to be consumed without need for further cycles of information gain or project wisdom, and it is certainly open to debate; panelists understand both viewpoints and will share their convictions and welcome audience members to do the same for a highly interactive discussion during this panel session.

**Expected Discussion Areas** Topics for discussion include but will not be limited to: Annotator and Team Member Training Can senior annotators train new annotators without having them go through the cross-training experience? What are best practices for annotator training? Are there NLP tools to facilitate the training? Has anyone ever tested the cognitive limits of the complexity of annotation tasks? How well do team members need to understand the science behind each aim? Evaluation Metrics: What evaluation metrics are available for measuring inter-annotator agreement? Are there any “out of the box” tools for calculating rates of agreement? Variant levels of project operational interpenetration: Are there ways to weight annotations according to whether the annotation class represents a clinical variable as opposed to a concept to be filtered? Should such considerations enter into the construction of annotation schemas? Scientific reproducibility: To what use can one put annotated data once the targeted NLP system has been trained and evaluated? Is there a way to use annotation data in conjunction with clinical annotation schemas to facilitate creating common data models? Interdisciplinary Contribution: Should annotators participate in project management and research team meetings? How do the strengths, weaknesses, and opportunities compare between the traditional and the interdisciplinary approach?

**Structure & Discussion Topics** The moderator will initiate the panel discussion, delivering a 10 minute summation of what motivated convening the panel, including a short description of the four key roles. Each of panel presentations (4 panels) of 10-12 minutes will address the tasks associated with their respective roles. Panelists will present workflows, and barriers to task completion as well as lessons learned, inclusive of cross-training experienced or received from other team members. The final 30 minutes will be interactive audience participation.

**Description of Panelists & Presentations**

**Panel 1 Team: Nancy Gentry, RN, MSN** is a Research Nurse Specialist, Vanderbilt University and at the Tennessee Valley Health System, VA. She has had experience in Staff Nursing in the Operating Room and clinical areas at several hospitals, and in Quality Assurance Nursing performing chart reviews of surgical patients at TVHS with the VASQIP program. Elizabeth Hanchrow, RN, MSN is a Research Nurse Specialist, Vanderbilt University and at the Tennessee Valley Health System, VA. She has been a Registered Nurse for 29 years and worked as a staff nurse in a variety of clinical areas including Medical Surgical, Cardiology, Oncology, Pediatrics, Urology, and Women’s Health, and in nursing informatics at Vanderbilt University Medical Center she supported computerized physician order entry and informatics research.

**Panel 1: Nancy Gentry, RN, MSN** together with Elizabeth Hanchrow, RN, MSN will present the role of the nurse annotator, and team function. **Planned discussion summary:** A team presentation of annotation protocol for training and error oversight will be followed by describing the factors contributing to their evolving role as collaborative team members. They will present how they leveraged the task of piloting and iterating annotation guidelines to bring the group to the awareness of the need to operationalize the tension between capturing clinical truth and delivering information to an NLP system. If there is time during the open discussion period they will demonstrate a graphic illustrating the operational procedures for annotation workflow which resulted from this group understanding.

**Panel 2: Brett R South, MS, PhD** is a Senior Scientist in the Department of Biomedical Informatics, University of Utah. Previously he was a Senior NLP Research Engineer for the Nuance Clinical Language Understanding group where he helped lead a group of 75 clinical language analysts tasked with large-scale semantic annotation of clinical corpora to support development of a computer-assisted coding module. His research interests include: clinical NLP, integrating efficiencies with manual human review tasks via improvements in tools, workflow modifications, or distributed review, human cognition, and data analysis. He will discuss the role of the data quality organizer in building a reference standard. **Planned discussion summary:** Dr. South’s presentation will be leverage his experience as data quality manager for many annotation projects including i2b2 2010/2011, de-identification,
disease surveillance, extracting information representing variables for carotid stenosis, readmission risk factors including social determinants of health (SDOH), and activities of daily living (ADLs)/instrumental activities of daily living (IADLs), functional status indicators, and other data normalization tasks. His presentation will be in the context of the data quality organizer focusing on the stages most common to NLP projects that are necessary to annotate clinical texts and generate reference standards that can be used for NLP system development and statistical performance evaluation. These common stages require independence in human review and require considerable effort and are often cognitively intensive. These stages include: 1) developing and pilot testing annotation guidelines and schemata; 2) training annotators; 3) performing the annotation task; 4) adjudicating disagreements and/or building consensus sets to create a final reference standard; and 5) estimating task reliability and validity. These tasks may also require clinical expertise at the sub-specialty level to reduce ambiguity in annotation guidelines and schemata, and to adjudicate disagreements. This presentation will discuss each of these stages as they relate to building reference standards that are of high enough quality to be used for NLP system development.

Panel 3: Glenn T. Gobbel, DVM, PhD, MS is a Research Assistant Professor in Health Services Research at the Tennessee Valley Health System and within the Department of Biomedical Informatics at Vanderbilt University. His work has focused on the development of near-real time NLP tools for extraction of medical information from free text and machine-learning based tools that rely on online learning for assisted annotation. He is currently a co-investigator on six nationally-funded projects focusing on the development of NLP systems for clinical use. His presentation will focus on the goals of annotation from the perspective of an NLP developer. Planned discussion summary: Dr. Gobbel’s presentation will be informed by his own experience developing NLP systems for extracting information with respect to cirrhosis, acute kidney injury, congestive heart failure, and patients undergoing cardiac stress tests. A key goal of annotation for an NLP developer is generally to create the data that can be used to produce, train, and test an NLP system to effectively extract medical concepts of interest and associated data. This presentation will describe some of the challenges associated with and approaches to generating such a system and creating the annotation schema and guidelines to accomplish this.

Panel 4: Steve Bradley, MD, MPH is Assistant Professor of Medicine in the Division of Cardiology at the University of Colorado, School of Medicine and a VA Health Services Research & Development Career Development Award (CDA) recipient. His research is focused on understanding the determinants of high-quality and high-value care delivery. He has conducted several national studies evaluating variation in the use and appropriateness of coronary angiography and PCI in addition to studies on the processes and outcomes of cardiovascular care delivery. In addition, he is the PI of a project grant funded by the VA to develop and validate an NLP application to abstract structured data from text reports of stress tests performed in the VA. This tool intends to address the absence of interpretable data on stress test results that limits opportunities to measure and support quality of risk factor modification and the high-quality use of procedural care for patients with ischemic heart disease.

He will discuss the role of the clinical expert in charge of the study design of the informatics project which consumes the products of NLP output. Planned discussion summary: As a cardiologist and health services researcher, Dr. Bradley seeks to understand the determinants of high-quality and high-value cardiovascular care. Currently, cardiovascular care lacks structured data on the indications and results of cardiovascular stress tests. This lack of data severely limits the ability to assess the quality of care for patients with ischemic heart disease. Accordingly, Dr. Bradley’s VA funded project sought to develop and validate an NLP application to abstract structured data from text reports of stress tests performed in the VA. Prior to working on this project, Dr. Bradley had no experience in the development of a NLP tool. In this panel discussion, he will share his experience in the transition from novice to somewhat experienced NLP tool developer from the perspective of a clinical champion who seeks to improve the quality of health care.

Ruth Reeves, PhD is a Health Services Research Scientist at the Tennessee Valley Health System, US Department of Veterans Affairs and an Assistant Professor in the Department of Biomedical Informatics at Vanderbilt University. She is currently principal investigator on a nationally funded merit award which deploys NLP methodology to temporal reasoning within the electronic health records of veterans suffering Post Traumatic Stress Disorder and is a co-investigator on six nationally-funded projects focusing on the development of NLP systems for clinical use. She will serve as the moderator of the panel. She will initiate the panel discussion, facilitate discussion and introduce each panelist with a brief statement of the role he or she represents within the annotation workflow. Planned discussion summary: Every research team member represents a cost to the project; no cost need be judged more burdensome than another if all knowledge and expertise represented within the research team can be optimized through continual and iterative cross-education among team members. Annotators bring the data to life and expose the complexities of both the project goals and the realities the study is meant to uncover. Developing and maintaining this touchstone of evidence is the lifeblood of successful clinical research.

Participation Statement: All proposed panelists have agreed to participate in the panel.
Interactive Panel: Open Architecture for Pathways/Care Coordination

Robert A. Greenes, M.D., Ph. D.\textsuperscript{1,2}, Steven J. Demuth, B.A.\textsuperscript{3}, Matthew M. Burton, M.D.\textsuperscript{3}, Davide Sottara, Ph. D.\textsuperscript{1,3,4}, Keith M. Toussaint, B.S.E.E.\textsuperscript{4}

\textsuperscript{1}Arizona State University Department of Biomedical Informatics, Scottsdale, AZ; \textsuperscript{2}Mayo Clinic, Scottsdale, AZ; \textsuperscript{3}Mayo Clinic, Rochester, MN; \textsuperscript{4}Healthcare Services Platform Consortium, Scottsdale, AZ

Abstract

Current health IT has evolved over many decades as a patchwork of proprietary legacy systems for managing electronic health records (EHRs), providing decision support, facilitating workflow, finances, and quality management, among other functionality. The closed, proprietary nature of the architecture and underlying technologies in these systems has resulted in silos of data, fragmentation of the patient records, difficulty coordinating care, and difficulty evolving to address the needs of the future. In order to better support the needs of clinical pathway development and care coordination, leading health care organizations should work together with leading open technology providers to create open architecture, open interface and open source community cloud computing environment. The presence of such an environment would enable a new generation of health information technology innovation resulting in information technology tools that would be better suited to meet the needs of a health care future where delivering best practice care to patients in every setting is the norm.

Intended Audience

The intended audience includes those who have an interest in applying modern, web-based architectural principles to health care information technology. More specifically, those who have a strong interest (pro or con) in introducing open architecture, open interface, open source and cloud-based approaches in health IT. The aim of the discussion will be to propose a broad industry effort to bring the aforementioned principles into the mainstream of health information technology development.

Introduction

This panel discusses opportunities, challenges, and possible approaches to creation of a new information technology ecosystem for health and health care. Much has been written about the forces driving the need for more connected, integrated, coordinated health and health care, the payment system reforms seeking to move health care in this direction, and the technology requirements for interoperable, standards-based architectures needed to bring it about.

Current health IT has evolved over many decades, particularly in the US, as a patchwork of proprietary legacy systems for managing electronic health records (EHRs), providing decision support, facilitating workflow, finances, and quality management, among other functionality. But these have been focused on particular hospitals, practices, or enterprises, resulting in silos of data, fragmentation of the patient records, difficulty coordinating care, and difficulty evolving to address the needs of the future. These needs include better summarization and visualization of data especially for complex patients, management of the patient’s care across multiple venues from home to office to hospital, more direct self-management of health and of care by patients, more comprehensive workflow management and decision support both for patients and providers, and the use of the ever-expanding volume of data to provide better analytic support for care decisions, quality management, research, and public health, and for other purposes, all in a manner that provides security and protects patient privacy.

Aim of the Discussion

The panelists represent some activities that are being taken to help bring about a new framework for interoperable apps and standards-based services to meet new health IT challenges. By briefly introducing them and describing the vision of how these relate to one another, we seek to stimulate interaction with the audience about the opportunities, obstacles, alternative approaches, and ways of proceeding that can lead to realization of the vision of a new vibrant ecosystem.
Specific Contributions of Each Speaker

Robert A. Greenes: The Arizona State University (ASU) Department of Biomedical Informatics is seeking to promote the further development of an evolving standards-based platform on which to develop apps and services that address needs of collaborating entities to meet evolving health care tasks not met by their EHR systems alone. The goal is to provide a sandbox and testbed development environment for evolving solutions that are interoperable and interface with underlying EHRs and other data sources and obtain and coordinate various services with those systems. Figure 1 is an example of a software architecture based on this approach. It provides opportunities to health care organizations to obtain add-on functionality that has been heretofore locked into proprietary vendor offerings, allows competition at the app level, and provides functions that bridge or extend those existing systems through evolving middleware services. An app design and deployment tool known as AppWorks is used to facilitate composition of multi-function apps using knowledge-based filters and event triggers.

Matthew M. Burton: In parallel and working with ASU colleagues, Mayo Clinic has developed software to support the optimal use of clinical pathways in the care of its surgical patients. The EASE (Enhanced Analytics for Achieving Surgical Excellence) project has built a care pathway management and monitoring application and framework that employs user interfaces that embrace clinician mental models, embodies the design principles of separation of concerns, uses open architectural principles (see Figure 2), open interfaces and open source components.

Jonathan R. Nebeker: The Veterans Administration’s Vista Evolution Program has an architecture and framework that supports interoperability, based on a three-tiered architecture that has similarities to those described above. Driving needs include care coordination: patient-centric, team-based care with capabilities for quality improvement; new user experience with multifaceted support for understanding and decisions that speed use of the EHR and improve quality of clinical reasoning; and capability for technical, semantic and process interoperability with external systems with which VA health care must communicate.

Oscar Diaz: The Healthcare Services Platform Consortium (HSPC) is a not-for-profit community of healthcare providers, software vendors, educational institutions and individual contributors who facilitate clinical application interoperability and data sharing by defining open, standards based specifications for enterprise clinical services and clinical applications. HSPC ultimately hopes to achieve interoperability and sharing of clinic assets across the entire health care domain.

Figure 1. An Interoperable-App-Enabling Architecture for Health Care, showing its 3-tiered architecture with apps, middleware services, and data sources at the 3 tiers.

Figure 2. EASE Solution Architecture
Keith Toussaint (panel organizer and moderator): Mayo Clinic and other health care organizations are seeking to adapt their business models and health care services to the transformation of the health care system and to continue to do what they do best while finding new ways to thrive and grow. The need for development of new health IT approaches is essential, and Mayo Clinic has been active in pursuing approaches to this challenge.

**Expected Discussion**

*Fostering a New Ecosystem for Health IT: Three Capabilities Needed*

A new ecosystem is in its very early stages for health IT, as a growing number of health care organizations (HCOs) are finding the need for it, as well as investing in it with their own internal funds and venture initiatives. The Federal government and many payers, health care researchers, and health professionals are trying to foster it. For such a vision to come about, three inter-related capabilities are proposed to be considered by the audience:

1. A specification for a standards-based architecture/framework that will foster interoperable apps and services – that is agreed upon and supported by both health care organizations, payers, vendors, and other stakeholders.

2. Methods for reliably operationally deploying apps and services in a robust, secure, mode – that ensures their ability to interoperate with and utilize existing EHR systems and other data sources, and to extend them in a smooth and highly reliable way. A leading approach to doing this is through a cloud-based Platform-as-a-Service ("PaaS") model.

3. Communally supported collaborative development resources – that enable (a) evolution of the capabilities of the standards-based architecture; (b) development and refinement of apps and services that use it; and (c) a migration path to reliable operational deployment and/or commercialization for apps and services that are developed and found to be suitable.

**Why Timely and Controversial**

Based on the initial success of the efforts by the presenters and interaction with colleagues in other institutions, we believe that there is a broad need for these capabilities across the health care sector. We therefore encourage an industry-wide effort to establish an open architecture, open interface, open source cloud-based initiative to facilitate interoperable apps for the advanced functionality which will enable further development of optimal care pathway and care coordination tools.

Nonetheless, such a framework does not currently exist yet in health care, and many factors such as legacy infrastructure, regulations, payment reform, and Meaningful Use requirements can greatly alter the speed and priority of adoption of interoperability and the forms it will take. There are alternative ways to achieve many of the goals, and heavy financial investment by both vendors and health care organizations and payers in them. There are also a number of organizational entities that are seeking to foster various aspects of interoperability. The feedback we are seeking should address the challenges that need to be faced and strategies that would be best to pursue.

**Agreement of Panelists**

All participants have agreed to take part on the panel.
(Authoring) Rules, (Distributed Query) Tools, and Drools: The challenging new world of high throughput phenotyping.

Jennifer A. Pacheco1; Abel Kho MD, MS1; Jyotishman Pathak, PhD2; Joshua C. Denny, MD, MS3; Shawn Murphy, MD4
1 Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; 2Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; 3Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA; 4Center for Research IS and Computing, Partners Healthcare, Boston, MA

Abstract
Deriving phenotypes from electronic health records (EHRs) is in progress within several national networks, including the electronic Medical Records and Genomics (eMERGE) network1,2, the Pharmacogenomics Research Network (PGRN)3, the SHARPn (Strategic Health IT Advanced Research Projects)4, and PCORnet5, to name only a few, and those phenotype definitions are also now being shared across those networks. Two main obstacles make it more difficult to share those EHR derived phenotypes across sites, and especially networks: 1) the absence of clear, unambiguous phenotype algorithm definitions represented in a standardized, if not executable, format that can easily be shared, and 2) the use of different standard data representations, and varying data models. This panel will discuss the varying array of methodologies, informatics tools, and data standards and data models for sharing and executing EHR derived phenotypes. The panel will also discuss efforts to harmonize data models and phenotyping informatics tools across the multiple parallel national efforts mentioned here.

A general description of the panel and issues that will be examined
Networks such as the Electronic Medical Records and Genomics (eMERGE) network have pioneered use of phenotype algorithms for use in EMRs, though their use in clinical studies predates this. Typically, phenotype algorithms have employed Boolean or machine learning methods across multiple classes of information (billing codes, medication data, natural language processing, and laboratory and test data) to accurately define cases and controls from EHR data. To date, eMERGE has investigated ~30 validated phenotypes via genome-wide association studies. In addition, eMERGE sites with other networks such as PCORnet have explored use of computable phenotype algorithms using tools such as KNIME (Konstanz Information Miner)6,7 and Drools to represent phenotype algorithms using computable languages with some success. Algorithms in eMERGE, PGRN, NIH collaborator, and PCORnet are being shared on PheKB.org.

The approach taken in i2b2 to date has been to use a statistical assessment; fitting a selection of possibly relevant attributes that describe a phenotype to a regression-type algorithm. This has been shown to accommodate to a wide variety of attributes that may be present or derived from the EHR, and is reasonably consistent form one site to another.8-9

New software tools are also currently being developed to author standardized phenotype algorithm representations that can be converted into executable format, such as one being developed by the Phenotype Execution Modeling Architecture (PhEMA) project10, which will use another common data model, the Quality Data Model (QDM)11,12. In addition, existing sharing of phenotype algorithms is usually in simple non-standardized textual and graphical format, let alone executable format, except in a few cases that we know of within the eMERGE network where executable workflows created using the open-source KNIME analytics platform are being used and shared by some eMERGE sites.

The panel will introduce and discuss with audience members involved, and/or interested, in sharing EHR-derived phenotypes, the following:

- Is a standard format needed for sharing phenotype algorithm definitions? If so, what should that look like? If not, why not? Are there several different formats we should be using depending on different use cases, and if so, what are those use cases?
- Should phenotype algorithms be shared ideally in executable format? If so, what should that format look like, & should it depend on a common data model? Are there different types of executable formats that
should be shared (i.e., KNIME vs. JBoss Drools), that would depend on the use case for those phenotypes? What are those use cases and thus what should those formats look like?

- Standard languages (and corresponding data representations) have been created for quality measures, such as the Quality Data Model. Are these sufficient for research uses, and why or why not? Or are common data models necessary for some use cases but not for others, & if so, which ones? How can we harmonize data (and data models) to enable more efficient and accurate sharing of phenotype algorithms? In addition, what data standards should be used?

- What are the most promising approaches to obtaining an accurate and portable phenotype? Rule based approaches and statistical approaches have been used to refine the phenotypes of the EHR which have afforded increases in sensitivity and positive predictive values. However there are tradeoffs of complexity, reproducibility, and portability which need to be made. Different approaches have optimized various parts of the desired outcome of obtaining an accurate phenotype.

Panelist’s perspectives

Jennifer A. Pacheco – Moderator

Mrs. Pacheco is a biomedical informaticist and the informatics lead for NUgene, the EHR-linked biobank at Northwestern University, and an active member of the phenotype working group of eMERGE, as well as PhEMA. She is also a previous member of PGRN. In these roles Mrs. Pacheco has co-led the development of over a dozen phenotype algorithms & led the validation and/or execution of over 50 EHR-derived phenotypes.

Abel Kho

Dr. Kho is an Assistant Professor of Medicine and Preventive Medicine at Northwestern University and a member of the phenotype working group of eMERGE and Co-Chair of the PCORnet Data Standards, Security and Network Infrastructure Task Force. He led Northwestern’s GWAS of genomic variants associated with Type 2 Diabetes (T2D) and more recently diverticulosis and diverticulitis. Dr. Kho will compare phenotyping experiences in eMERGE and PCORnet.

Jyotishman Pathak

Dr. Pathak is an Associate Professor of Biomedical Informatics at the Mayo Clinic College of Medicine. His research focuses on developing methodology and applications for standardized and computable phenotyping algorithms. Dr. Pathak will describe recent progress from the SHARPn collaboration developing open-source solutions for high-throughput and scalable phenotyping using Hadoop-based big data technologies, and highlight ongoing work in creating a robust platform for phenotyping algorithm authoring and dissemination. Dr. Pathak has also led the development of phenotype data harmonization platform within eMERGE and PGRN.

Joshua Denny

Dr. Denny is an Associate Professor in the Department of Biomedical Informatics at Vanderbilt University, and was the SPC chair of the 2014 Summit on TBI. Vanderbilt uses an opt-out model biobank associated with a deidentified Synthetic Derivative of the EHR. Vanderbilt’s has used BioVU for more than 100 studies, including investigation of >40 diseases and >30 pharmacogenetic phenotypes led by Dr. Denny. Many of these algorithms have been originated in or have been shared with the eMERGE or Pharmacogenomics in Very Large Populations (PGPop) networks. For example, one study selected individuals without cardiac disease and studied influences on cardiovascular traits, finding variants influence slower intracardiac electrical conduction. Phenome-wide association study (PheWAS) demonstrated some of these variants confer risk for atrial fibrillation, and a survival analysis of this original heart healthy population demonstrated similar risk of development of atrial fibrillation during the up to 20-year follow-up in the EHR. He also developed algorithms and software to perform PheWAS, scanning the EHR phenotype for genetic associations with many diseases not anticipated in the initial GWAS, and use of this tool to explore pleiotropy in current GWA studies. Dr. Denny has also led development of PheKB.org to share phenotype algorithms and phewascatalog.org to share PheWAS results.

Shawn Murphy

Will introduce i2b2 and its federated query platform SHRINE (Shared Health Research Information NEtwork) as a research informatics platform and their position statement including approaches it has taken to solve problems in phenotyping using the EHR. Dr. Murphy will continue with discussion of i2b2 design philosophy as it regards data
modeling, ontologies, standards and terminology systems and how these can be woven together to provide the basis for dealing with the many aspect of phenotyping in our community. Will describe the optimal i2b2 ecosystem (tools, application development frameworks it supports, products and use-cases) and end with the future development roadmap for providing phenotyping tools.

**Plan for interaction between panelists and the audience**

The moderator will encourage the audience to interact with the panelists through questions such as those above and by encouraging discussion and comments on current methods for sharing and executing phenotypes within and across the aforementioned national initiatives.

All participants have agreed to take part on the panel.

**References**

3. Pharmacogenomics Research Network, pgm.org
Abstract

Many healthcare institutions face a second crisis in electronic health record system (EHR) implementation: migrating from one mature commercial or home-grown EHR to another. A considerable portion of the healthcare market is now converting to a second or third EHR system. AMIA members, ostensibly, have examined in great detail the challenges of initial EHR selection, implementation, adoption, and the shift away from paper records. Research is missing or sparse to guide informaticists and information technology specialists on the challenges and support for subsequent implementations.

This panel seeks to accomplish the following:

- Recognize that one important implementation issue is migrating from one mature EHR to another;
- Explore the advantages of such shifts;
- Discuss extant evidence-based literature on system migration, or “re-implementation”;
- Identify the conversion risks and costs;
- Distinguish often unforeseen consequences;
- Guide others in this common endeavor.

Panelists will explore what is known via anecdotal sources, listserv discussions, personal experience, and the relatively few published studies related to this topic. We shall solicit information from fellow AMIA members and present new information. Panelists will declare their positions on these topics, followed by questions proffered to each other in an attempt to stimulate audience participation in the open discussion period.

General Description

Description of intended audience

The intended audience consists of the following people: clinicians in institutions considering or undergoing EHR migrations; frontline operations informatics professionals, including Chief Medical Informatics Officers, as well as information technology leaders such as Chief Information Officers and Chief Operating Officers, who frequently manage implementations and transformations in healthcare institutions, trainers, HR professionals who must adjust workloads during the training and ramp up process, Chief Financial Officers who must manage the costs of these shifts, patient relations personnel who are often called upon to smooth transitions and not alarm patients as clinicians learn to use the new systems; as well as informatics researchers who examine clinical information system and
implementation issues. These clinical leaders and researchers need the input of users of systems who are impacted by such a change, as usability is a critical factor that can predict the success or failure of such a transformation.

**Issues that the panel will address**
- Presentation of the new paradigm towards system migration: Why break what’s not broken? Or isn’t it?
- What is driving this new paradigm: consolidation, mergers, acquisitions, affiliations, and new payment modalities?
- What does the community hospital need to do?
- What does the academic medical center need to do?
- What do Critical Access Hospitals and other small, rural hospitals need to do?
- What are the human, sociological, and financial costs of these transitions?
- What are the risks?
- Can smaller institutions keep up with the financial strain of conversions?
- What role do larger institutions have in terms of population health, accountable care, and value based payments?

**Aims of the discussion**
The goals of this interactive panel discussion are to:
- Introduce members of AMIA and the informatics community to this emerging—but increasingly common—phenomenon;
- Propose and discuss expected challenges when changing EHR products;
- Develop the appropriate research questions to enable objective investigation of the issues

**Specific contribution of each speaker**
The panelists will each speak for approximately 10 minutes. The perspective of the first speaker is to introduce the paradigm shift, and present the point of view of the community hospital. The second speaker will address the perspective of the large academic or university center. The third speaker will discuss the position of Critical Access Hospitals and other small, rural hospitals and will assess the risks for those clinicians, clinics, and hospitals that may not be able to participate in such a transformation. The fourth speaker will review the human, sociological, and financial impacts of EHR shifts. All panelists will then answer questions addressed to them from the other panelists, and from the audience.

**Expected discussion**
Each speaker will pose provocative questions to at least one of the other speakers on the panel, and each speaker will be given 5 minutes to answer in the format of a friendly debate. The audience will have a chance to participate in the remaining 30 minutes of the program.

**Explanation of why the topic is timely and controversial**
Interoperability, Meaningful Use, optimization of clinical decision support, and payment system reform have dominated recent discussions regarding EHRs. Part of AMIA’s mission is to keep informaticists and related scientists informed about, and prepared to address, emerging trends in healthcare information systems. Concern about the consolidation of the vendor marketplace will add additional urgency to these phenomena. We already observe examples of problems when healthcare entities merge, but their EHR systems do not. This panel will explore all aspects of this new problem.
Panel Members

Richard Schreiber, MD

Richard Schreiber, MD, FACP, is Chief Medical Informatics Officer at Holy Spirit Hospital (HSH)—a Geisinger Affiliate—in Camp Hill, PA, who is board certified in Clinical Informatics, and is also a general internal medicine hospitalist.

John D. McGreevey III, MD, FACP

Dr. John D. McGreevey III is a general internal medicine hospitalist attending at Penn Medicine, an Assistant Professor of Clinical Medicine at the Perelman School of Medicine at the University of Pennsylvania, and a general internal medicine hospitalist.

Catherine K. Craven, PhD, MLS, MA

Catherine K. Craven, PhD, MLS, MA, Immediate Past Chair, Evaluation WG, graduated from the doctoral program in Health (clinical) Informatics in December 2014, from University of Missouri, Columbia, Mo. Her research focuses on rural and critical access hospitals.

Ross Koppel, PhD, FACMI

Dr. Koppel studies, and is an advocate for, improving healthcare IT—the implementation process, workflow, human-computer interactions, vendor-provider relations, legal and ethical aspects of CDS and HIT, and of HIT’s usability. His articles in JAMA, JAMIA, Annals of Internal Medicine, NEJM, Health Affairs, J of Patient Safety, etc. are often considered seminal works because his original training as a sociologist of work, the professions and technology allows him to combine ethnographic research, extensive statistical analysis, surveys, observations, interviews, organizational and financial analyses, and usability studies. Dr. Koppel is a Fellow of the American College of Medical Informatics.

Participation Statement

I, Richard Schreiber, hereby confirm that all panelists listed in this proposal have agreed to participate in this panel. Panelists are aware that there are no travel funds available. Panelists are also aware that the Working Groups are unable to reimburse their registration costs.

References


Patient privacy and “de-identified” health records in the genomic era

Jessica D. Tenenbaum, PhD1, Greg Biggers2, Leslie E. Wolf3, Bradley Malin, PhD4, Lucila Ohno-Machado, PhD5

1Duke University, Durham, NC 2Genomera, Palo Alto, CA 3Georgia State University College of Law, Atlanta, GA 4Vanderbilt University, Nashville, TN 5UC San Diego, La Jolla, CA

Abstract

Current policy allows researchers to use de-identified data from electronic health records for research purposes. Data is considered de-identified if eighteen identifying fields, including name, address, phone number, social security number, etc. are stripped off of the derived dataset. However, in the era of precision medicine, detailed genomic data will increasingly be stored along with patients’ clinical data. Genomic data, particularly whole genome/exome sequencing is, by definition, identifying. The availability of genomic data together with clinical information represents a rich resource for science, but also introduces ethical challenges around patient privacy. How can we manage the research demand for data while respecting patients' concerns, interests, and rights? On what "truths" does everyone agree today, and what are the potential sticking points down the road? What can we, as informatics researchers, do to help? These are just some of the timely and compelling issues that will be addressed by this diverse panel, featuring technical, ethical/legal, community, and patient’s rights perspectives.

Introduction

Current policy for secondary use of health records allows researchers to use electronic health record (EHR) data for research, so long as the records are “de-identified” by stripping out 18 identifying fields, including name, address, phone number, social security number, etc. However, in the era of precision medicine, detailed genomic data will increasingly be stored along with people’s clinical data. The availability of genomic data together with clinical information represents a rich resource for science, but also introduces ethical challenges around patient privacy [1].

There are some who argue that the risks of re-identification have been overstated, causing unnecessary obstacles toward scientific progress. Others have demonstrated the relative ease with which they were able to re-identify an individual based on publicly available genomic and ancestral information. Up until now, these issues have been largely academic, but as more genomic data is incorporated into health records, and as we move toward a future in which individuals have their genomes sequences as part of routine care, the issue becomes more germane. The advantages to using such data for research will be significant, particularly in light of the large sample size required for statistical significance for any but the most straightforward genotype/phenotype relationship. However, if we continue on the current path, and with the current interpretation of the law, patients who have never consented to research will have their unique genomes shared on servers across the world without ever being aware.

It should be noted that these issues have implications not only for secondary use of EHR data, but for medical research as well. If DNA is designated as identified data, then the possibility of studying it as part of a limited or de-identified dataset is eliminated. Similarly, NIH Genome Sequence Data Sharing policy currently states that DNA is de-identified, but it acknowledges there is a potential for identification, such that it requires consent for every record submitted to dbGaP moving forward.

Researchers, together with other key stakeholders including patients and patient advocates, must identify and discuss these issues sooner than later, so that all parties can understand the issues and perspectives at hand, and come to agreement as to how to proceed moving forward.

Timeliness of Topic

This topic is particularly timely in light of the White House’s recent announcement around the Precision Medicine Initiative, aimed at uncovering “individual differences in people’s genes, environments, and lifestyles” to enable more targeted and effective treatments. [2]
Intended Audience
The target audience for this panel is quite broad, ranging from investigators whose research uses electronic health record data, to those who focus on genetic and genomic data for whom this new landscape will be particularly alluring, to anyone who has, or may someday have, genetic data associated with his or her medical records.

Speaker Contributions
This panel will discuss the issues raised above from several different perspectives: ethical/legal (Wolf), technical (Ohno-Machado, Malin), community (Malin, Biggers) and patient perspectives (Biggers).

Jessica D. Tenenbaum- Duke University- Moderator
Dr. Tenenbaum will frame the topics to be addressed and introduce the panel speakers, and will moderate the presentations and the discussion to follow.

Leslie E. Wolf- Georgia State University School of Law
The coming existence of millions of interoperable electronic health records with genomic sequence or SNP data will present researchers with tremendous opportunities. But using those records for research without consent (as permitted under existing laws) risks undermining public support for research. How can we manage the research demand for data while respecting patients’ concerns, interests, and rights?

Lucila Ohno-Machado- UC San Diego
Dr. Ohno-Machado will discuss patient preference elicitation for sharing clinical data and biospecimens. She will provide preliminary results of a study being conducted at two academic medical center clinics in which patients decide what to share and the CTSA informatics team delivers data to investigators as appropriate. She will also discuss how genomic data could be accessed for research without compromise of patient privacy, and the security environment in which the data should reside.

Brad Malin- Vanderbilt University
Dr. Malin will discuss the role of community consultation and advisory groups as it pertains to opt-in and opt-out. He will also discuss how secondary use of genomic and clinical data post-consent (regardless of the opt-* strategy) relate to the notions of de-identification and the extent to which this is a worthwhile protection given current threats to such protections. His comments will be centered around his experience with the establishment and management of a large biorepository at Vanderbilt University and other medical centers involved in the NIH-sponsored Electronic Medical Records and Genomics (eMERGE) Network.

Greg Biggers- Genomera, Genetic Alliance
How do we navigate the tensions between vigorous biocivics and the emerging reality that de-identification of precision medicine data is a myth? How do people act when privacy and health seem to be at odds with one another? And what happens when subjects become researchers? Biggers will discuss lessons in privacy, control, engagement, and trust learned from grassroots projects including Genomera (participant-driven health studies) and Genetic Alliance's PEER (Platform for Engaging Everyone Responsibly), as well as his own experience as a patient with a sequenced genome.

Panel Discussion Aims and Expectations
We expect a lively discussion to follow the presenters’ prepared comments. The topic is not highly technical, but is one that could, and most likely will, someday affect almost every conference attendee. Discussion is likely to address patients’ rights to privacy, whether we enter into an implicit social contract when we seek medical attention, issues of probability and numeracy, legal protections (both existing and desired), appropriate consequences of data abuse, and more.

Conclusion
Data privacy in the genomic era is a complex but addressable issue. As it stands, only a small fraction of the population currently have genomic data associated with their medical records, but the numbers are increasing every day. The time to have these conversations is now.
Participants’ Agreement to Attend

All speakers have indicated the intention to participate in this panel.

References

Building the Computational Workforce for Precision Medicine

Jessica D. Tenenbaum, PhD¹, Joshua C. Denny, MD, MS², David B. Flannery, MD³, Douglas Frdisma, MD, PhD⁴, Marc S. Williams, MD⁵

¹Duke University, Durham, NC ²Vanderbilt University, Nashville, TN ³American College of Medical Genetics and Genomics, Bethesda, MD ⁴American Medical Informatics Association, Bethesda, MD ⁵Geisinger Health System, Danville, PA

Abstract

Biomedical research and practice have never been more interdisciplinary than they are today, and yet interdisciplinary training in biomedicine remains the exception and not the rule. AMIA has a number of initiatives, from working groups to the annual Summit on Translational Bioinformatics, which emphasize the importance of the “bio” end of the biomedical informatics spectrum. However, there is significant opportunity for collaboration with other organizations with complementary expertise. AMIA has significant capabilities in informatics to support precision medicine. The American College of Medical Genetics and Genomics has significant expertise in medical genetics and genomics to improve human health. This panel brings together thought leaders from these two organizations to discuss the challenges in training tomorrow’s workforce for precision medicine, and the possibilities for interprofessional educational opportunities.

Background

Biomedical research and practice have never been more interdisciplinary than they are today. This fact is underscored by a number of trends in recent years, not the least of which has been the Health Information Technology for Economic and Clinical Health (HITECH) Act and its requirements for the adoption and meaningful use of electronic health records. More recently, in January 2015, President Obama announced a $215M Precision Medicine Initiative aimed at the treatment and prevention of disease that take into account individual variability in genes, environment, and lifestyle. [1] Despite these initiatives and the significant need for interdisciplinary education, such training in biomedicine remains the exception and not the rule.

AMIA has had a long standing interest in, and commitment to, the importance of biology in the broader field of biomedical informatics. The theme of AMIA’s annual symposium in 2002, titled “Bio*Medical Informatics: One Discipline”, could be seen as a harbinger of themes from the decade plus since, including genomic, personalized, and precision medicine. Similarly, the establishment of the annual Summit on Translational Bioinformatics in 2008 served as a professional gathering for biomedical informaticists whose focus lies at the molecular end of the biomedical spectrum. Beneficial though that venue has been to this community, communication and cross pollination with other organizations focused in this area has been suboptimal. This proposal, submitted in collaboration with the American College of Medical Genetics and Genomics (ACMG) aims to enhance communication between our respective organizations and to spark ongoing collaboration through exploration of this topic of mutual importance, potentially leading to opportunities in interprofessional educational opportunities.

Introduction

The establishment of Clinical Informatics as a Board Certified medical specialty in 2011 was an important step toward training an informatics-enabled workforce of tomorrow, and 2013 and 2014 saw more than 700 physicians board certified in clinical informatics. AMIA played a key role in the establishment of this sub-specialty, through an 18 month, RWJF-funded process to define the core content of the subspecialty of clinical informatics and the training requirements for proposed clinical informatics fellowships. For professionals not eligible for certification in a medical sub-specialty, e.g. non-physicians, the Advanced Interprofessional Informatics Certification (AIIC), arose out of AMIA's Academic Forum in 2011. Directed by Don Detmer, MD, past president and CEO of AMIA, the AIIC aims to develop a qualified and sustainable informatics-trained workforce not only as a medical subspecialty, but also as a viable career path for non-clinicians in complementary fields, e.g. informaticians, computer scientists, dentists, public health professionals, etc.

The American College of Genetics and Genomics (ACMG) provides education and resources for professionals in the field of medical genetics. It also serves as the voice for these professionals on important topics of interest to the larger biomedical community. A prominent recent example is the ACMG’s recommendations regarding return of results in
clinical exome and genome sequencing. [2] The ACMG leads the development of professional practice guidelines for both clinicians and laboratorians, creates educational content (e.g. ACT sheets for newborn screening and other disorders) to aid non-geneticists in managing patients with rare genetic conditions. ACMG is also involved in the NHGRI funded Clinical Genome (ClinGen) project which will be developing online resources and informatics tools to support the use of genomic variant information in clinical care. [3] At present accredited genetic training programs do not have defined competencies, formal training in, or exposure to informatics. The leadership of the ACMG recognizes informatics as essential to the realization of Precision Medicine and is looking to partner with content experts such as AMIA to provide content for their trainees.

Key areas for enhanced training opportunities include (but are not limited to): genetics, genomics, high throughput technologies, biostatistics, molecular systems biology, data standards and analytics, clinical informatics, clinical decision support, evidence-based patient care, human factors engineering, health IT systems and applications, and team science.

In this panel, AMIA thought leaders in the areas of precision medicine have come together with colleagues from the ACMG to discuss the issues and opportunities that arise as we begin to lay out a strategy for training the workforce that will be required to execute on the vision of Precision Medicine, and how our organizations can work together to achieve these educational goals.

Intended Audience
This topic will be of interest to a broad AMIA audience including informatics professionals who are currently practicing or interested in pursuing training in genetics and genomics, as well as educators who aim to expand their department’s offerings, whether within their home institutions, or virtually online.

Topic Timeliness and Controversy
The federal government has announced its intention to devote $215 million in funding to the Precision Medicine Initiative in the President’s 2016 budget. More importantly, as discoveries are made through this initiative and translated into clinical practice, the distribution of the $3 trillion spent each year in the US will inevitably shift. Those trained to leverage and execute on these new discoveries will be at a significant competitive advantage.

On the other hand, the US biomedical enterprise a large ship to steer. Many stakeholders have much to lose, in dollars, stature, or otherwise, from a change in the status quo. Entrenched authorities in medical training and practice may raise opposition to any type of overhaul of education or implementation of medical practice.

Speaker Contributions
Jessica Tenenbaum, PhD- Duke University
Dr. Tenenbaum is Chair of AMIA's Genomics and Translational Bioinformatics Working Group. She will frame the panel, introduce the speakers, and moderate the presentations and the discussion that follows.

David Flannery, MD, FACMG- Medical Director, ACMG
Dr. Flannery will provide an overview of the American College of Medical Genetics and Genomics (ACMG) to the AMIA membership. He will discuss current training programs accredited by the American Board of Medical Genetics and Genomics and the need for informatics as part of the training curriculum. Dr. Flannery will present some of the ACMG materials that could be adapted for clinical decision support as well as clinical research activities such as the Newborn Screening Translation Research Network and the ClinGen project.

Josh Denny, MD, MS, FACMI- Vanderbilt University
Much research in the last 5 years has demonstrated the ability to use electronic health records (EHRs) for genomic analysis. A key result from these studies is that it is possible to extract research-quality phenotypes using EHR data, but typically this is an involved and iterative process between informaticians and clinical experts. Typically, these have been with diseases, though a few studies have shown success looking at drug-response traits. We now sit with the possibility to explore deeper and more granular phenotypes, dissecting disease into a more granular and precise phenome. Dr. Denny will describe the multidisciplinary infusion of clinical expertise that will be required for continued advancement in these areas, with training in data mining, natural language processing, computational as well as classical genetics.

Douglas Fridsma, MD, PhD, FACMI- President & CEO, AMIA
As AMIA’s President and CEO, Dr. Fridsma will discuss a number of AMIA training initiatives, including the Advanced Interprofessional Informatics Certification (AIIC), AMIA’s "10 x 10" courses, the Joint Summits on Translational Bioinformatics, and other AMIA initiatives. He will offer a high level perspective on how the various initiatives tie together, and how AMIA as an organization can work with the American College of Genetics and Genomics (ACMG) to train the Precision Medicine workforce of tomorrow.

Marc S. Williams, MD- Director Genomic Medicine Institute, Geisinger Health System

Dr. Williams is one of a handful of individuals boarded in clinical genetics and clinical informatics. In his presentation, Dr. Williams will briefly compare the competencies for each certification and propose content for each of the respective certification programs that could define interprofessional training. Opportunities to co-manage and harmonize Maintenance of Certification requirements to reduce the burden on individuals boarded in both specialties will be explored. He will also briefly explore informatics training to allied genetics professionals such as genetic counselors through the proposed AIIC program.

Questions for Discussion

- What are the key competencies to include in a clinical informatics program with a genomic emphasis?
  - What are the minimal genomic competencies to include in all clinical informatics programs?
- What strengths do current training programs have in genomics/precision medicine?
- What are the key gaps in current clinical informatics training that must be addressed?
- How can interactions between bioinformatics, genetics, and clinical informatics training programs be promoted to support precision medicine?
- What opportunities exist to introduce the multi-disciplinary nature of the field early in training, e.g. at the undergraduate level and professional training, such as medical school?
- How useful/important is an MD-PhD, really?

Conclusion

AMIA is poised to make significant contributions to the Precision Medicine Initiative. It also stands to benefit significantly from strategic partnerships with complementary professional organizations like the high profile, high impact ACMG. This panel will frame both the opportunities and challenges around collaboration and cooperative education. The broader discussion is sure to serve as a launching point for a productive partnership moving forward.

Statement of participation

All panelists have agreed to attend and participate if this panel is selected for the AMIA 2015 scientific program.

References

Opportunities for Social Media within Consumer Health Informatics

Rupa Valdez, PhD1; Sahiti Myneni, PhD, MSE2; Andrea Hartzler, PhD3; Lena Mamykina, PhD4; Nathan Cobb, MD5; Laura Barnes, PhD6 (moderator)

1Department of Public Health Sciences, University of Virginia, Charlottesville, VA; 2University of Texas School of Biomedical Informatics at Houston, Houston, TX; 3GroupHealth Research Institute, Seattle, WA; 4Department of Biomedical Informatics, Columbia University, New York, NY; 5Georgetown University Medical School, Washington, DC; 6Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA

Abstract
Social media, including blogs, wikis, discussion forums, and social networking tools, has become an increasingly significant space for health-related interactions among patients and others engaged in their care (e.g., family, friends, peers, and health professionals). This panel will address critical questions surrounding current and future opportunities for social media within consumer health informatics. Panelists will discuss their experiences analyzing health-related interactions on social media and the resulting implications for consumer health informatics design and research: Dr. Hartzler will discuss ethical considerations; Dr. Myneni will discuss use of existing theories and methods for data analysis; Dr. Valdez will discuss novel recruitment and data collection methods; Dr. Mamykina will discuss design guidance drawn from observational learning; Dr. Cobb will discuss methods of outcome measurement and inherent challenges. The session will then serve as a forum for panel and audience engagement with opportunities at the social media and consumer health informatics intersection, including: innovative research designs, strategies for safeguarding user privacy, novel methods, and the role for health professionals and other stakeholder groups in social media interactions. This panel is timely given that the rapid growth of social media for personal health management has enabled new and evolving forms of patient engagement and stakeholder collaboration.

Intended Audience
Individuals who have experience or interest in leveraging social media for consumer health informatics design and research.

Aim of the Discussion
To discuss the potential challenges, strategies, and ethical concerns associated with using social media within consumer health informatics design and research.

Overall Panel Topic
Currently, one in four Americans use social media to seek health information, while social media use in general continues to rise across all demographic groups. Although people use a range of social media for health-related purposes, two general forms are prevalent in health-related interactions: online health communities (e.g., PatientsLikeMe, CaringBridge, TuDiabetes) and general social media platforms (e.g., Facebook, Instagram, Pinterest), many of which have dedicated spaces for health-related interactions. The rapid growth of social media for personal health management has enabled new and evolving forms of patient engagement. As the emerging suite of platforms for personal health management continues to unfold, questions regarding the functions and effectiveness of social media for supporting patients provoke a multitude of responses. Therefore, the purpose of this panel is for panelists to discuss their experiences investigating health-related interactions on social media platforms and to converse with the audience about considerations and opportunities for future consumer health informatics design and research efforts.

Contribution of Each Panelist
At the start of the session, each panelist will provide a short presentation about his or her work at the intersection of social media and consumer health informatics.

Andrea Hartzler, PhD
Dr. Hartzler will focus on considerations for research ethics related to social media and the resulting implications for consumer health informatics design and research. She will differentiate types of social media by the style of collaboration and choices of research design each support, from large-scale text processing to small-scale interviews. Across research designs for studying online communities, Dr. Hartzler will contrast strategies for preserving user privacy and confidentiality in data collection and dissemination of results. By characterizing the variety of research options and ethical considerations that social media present, this presentation will facilitate discussion on understanding ethical challenges and opportunities for social media design and research in consumer health informatics.

*Sahiti Myneni, PhD, MSE*
Dr. Myneni will focus on the need to adopt and modify existing theories and methods to analyze social media data and the resulting implications for future consumer health informatics design and research. Specifically, she will address how data analysis approaches should account for scalability, granularity, and theoretical validation. The presentation will highlight the role of informatics, specifically automated text analysis techniques, in achieving balance between analytical depth and scale. Dr. Myneni will also discuss how application of behavioral change theories and social influence models may be used to examine consumer engagement in online health communities. By presenting these data analysis techniques, this presentation will facilitate discussion on how existing theories and methods may be adapted or modified to accelerate social media design and research in consumer health informatics.

*Rupa Valdez, PhD*
Dr. Valdez will focus on the opportunity for novel recruitment and data collection methods and the resulting implications for future consumer health informatics design and research. Dr. Valdez will address approaches to leveraging existing social structures embedded within social media (e.g., Facebook groups and pages) to recruit and engage participants in online versions of traditional data collection methods (e.g., interviews and surveys) and to understand participants’ perspectives on their social media-based health-related interactions. By illustrating novel recruitment and data collection methods, this presentation will encourage discussion on the need for methodological development to accelerate social media design and research in consumer health informatics.

*Lena Mamykina, PhD*
Dr. Mamykina will focus on opportunities to leverage social media to enable observational learning in the context of health and wellness through sharing of lived experiences. She will discuss recent studies of social and observational learning on such sites as Pinterest and Instagram, and will draw implications for the design and research of informatics interventions for enabling and facilitating these practices. By describing the potential for drawing design implications from social media-enabled observational learning, this presentation will stimulate discussion on translation of knowledge gained through field studies into design guidance.

*Nathan Cobb, MD*
Dr. Cobb will focus on how social features of real world, large-scale networks can lead to changes in a range of outcome measures and the resulting implications for future consumer health informatics design and research. Drawing from large scale randomized control trials conducted within online social networks, he will discuss how social processes can drive diffusion and mediate both engagement and health outcomes. He will specifically discuss evidence that diffusion of health interventions, ie viral spread, can be measured as an outcome as part of a design process of online interventions. By describing these methods this presentation will enable discussion of both traditional and alternative outcome designs and their ethical implications.

**Panel Topics**
After the introductory presentations, the panel and audience will engage in a discussion of various topics, facilitated by the moderator. Topics relevant to this panel are:

- What are the major ethical challenges to conducting social media-related consumer health informatics research and design?
- What are the major practical challenges to conducting social media-related consumer health informatics research and design?
- How do we promote research at the intersection of social media and consumer health informatics?
• How can we engage patients and other stakeholders in research and design at the intersection of social media and consumer health informatics?

• What theories can we draw on across disciplines to serve as foundations for research and design at the intersection of social media and consumer health informatics?

• What methodological innovations are necessitated by research at the intersection of social media and consumer health informatics?

• How do we effectively and efficiently translate understanding of patients’ and other stakeholders’ use of health-related social media into recommendations for consumer health informatics design?

• What are appropriate evaluation measures for consumer health informatics-related social media interventions?

After discussion of the topics above, the moderator will lead the panelists in a debate to address the following question: What are the priorities for future research and design at the intersection of social media and consumer health informatics?

Goals of the Panel Discussion
(1) Foster an interdisciplinary dialogue related to the opportunities for social media within consumer health informatics; (2) Provide AMIA symposium attendees with an understanding of the multiple ways in which they may leverage social media within consumer health informatics research and design practice; (3) Provide AMIA symposium attendees with an understanding of the current challenges and ethical dilemmas surrounding the use of social media within consumer health informatics design and research; (4) Engage AMIA symposium attendees in a collaborative discussion regarding appropriate future directions for work at the intersection of social media and consumer health informatics.

Panel Timeliness
This panel is timely because the dramatic rise in social media for personal health management has enabled new and evolving forms of patient engagement. Appropriately studying and designing to support this novel form of patient engagement requires consideration of unique ethical issues and the development of new theoretical, methodological, and translational approaches. Because of its interactive nature, this panel will serve as a forum in which audience members and panelists can collaboratively advance dialogue at the intersection of social media and consumer health informatics. Audience members will have the opportunity to learn from the panelists and one another about new developments at this intersection. The sharing of experiences stimulated by this panel will serve as a foundation for generating and prioritizing future design and research initiatives required to realize opportunities for social media in consumer health informatics.

All panelists have agreed to take part on this panel.
Health Informatics Graduate Program Accreditation: CAHIIM Process and Standards Update

Judith J. Warren, PhD, RN¹, Stephen Johnson, PhD², Suzanne Boren, PhD, MHA³, Stuart Speedie, PhD⁴, Guenter Tusch, PhD⁵

¹Warren Associates, LLC, Plattsmouth, NE, ²Cornell University, New York City, NY, ³University of Missouri School of Medicine, Columbia, MO, ⁴University of Minnesota, Minneapolis, MN, ⁵Grand Valley State University, Allendale, MI

Abstract
Academic program accreditation in higher education is both a process and a status involving a review of a program against formal Standards to determine educational quality, how well students and society are served, and if successful the award of accreditation. The Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM) has offered accreditation at the masters’ degree level in Health Informatics since 2010. AMIA officially joined CAHIIM as a Member organization in January 1, 2015. A revision in the governance structure created a Health Informatics Accreditation Council with the responsibility of establishing and maintaining accreditation Standards, policies and procedures. This Panel will present the Council’s work to date and address questions about how CAHIIM accreditation is conducted, the value of accreditation to the health informatics field, the status of goals for 2015-2016, how the process is conducted, and the types of CAHIIM volunteer opportunities. Participants will have the opportunity to ask questions and provide input on the Council’s initiatives from their varied perspectives.

Introduction
Academic program accreditation in higher education is both a process and a status involving a review of a program against formal Standards to determine educational quality, how well students and society are served, and if successful the award of accreditation. The Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM) has offered accreditation at the masters’ degree level in Health Informatics since 2010. CAHIIM is a recognized accreditation organization by the Council for Higher Education Accreditation (CHEA). Effective January 1, 2015, AMIA joined CAHIIM as an official Member organization of CAHIIM and a revision in the governance structure has provided new opportunities for volunteers to participate.

Aims of Discussion
Judith J. Warren, PhD, RN, FAAN, FACMI will introduce the Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM) structure and the Health Informatics Accreditation Council. Volunteers and staff supporting the CAHIIM organization from early discussions to the present, in the audience will be recognized including the CAHIIM Board of Directors. She will elaborate on the scope of CAHIIM Health Informatics accreditation, the volunteer appointment process and the criteria to serve on the Panel of Accreditation Reviewers who conduct the site visits and prepare summary reports to CAHIIM.

Stuart Speedie, PhD will provide a brief overview of the journey AMIA has taken to collaborate with CAHIIM, why the need for accreditation of graduate health informatics programs was determined, and the value proposition that comes from a defined academic program accreditation process. He will briefly remark on any revisions to the Standards for Accreditation of Health Informatics Master’s Degree Programs and the timeline for implementation, as a recent responsibility of the Health Informatics Accreditation Council.
Stephen Johnson, PhD, FACMI will address the Health Informatics Competencies Framework project, program outreach through recent webinars and comments to date, as well as the next steps and timeline to implementation. The impact of a set of core competencies to the health informatics professional community is a significant step in public recognition of this academic discipline.

Suzanne Boren, PhD, MHA will share practical experiences with her institution’s Health Informatics master’s degree program as it currently moves through the self-assessment phase of the accreditation process. She will share several practical tips on preparation and documentation requirements.

Guenter Tusch, PhD, will moderate the panel of presenters and field audience questions at the conclusion of the presentations. CAHIIM staff will be present to address any questions on CAHIIM fundamentals, policies, and procedures related to the Health Informatics accreditation process.

Significant time will be allocated to a question and answer period giving the audience an opportunity to reflect on the progress made, and to share ideas and perspectives on the health informatics accreditation process with the Council and staff. An added dimension will be an opportunity for audience comments on any proposed revisions to the 2010 Accreditation Standards for Health Informatics Masters Degree Programs and the Competencies Framework that will have been made available to the AMIA membership in advance of the November conference.

**Conclusion**

Accreditation of health informatics graduate programs is essential to ensure a supply of highly qualified professionals in the healthcare field able to address the growing needs of information technology applied to all realms of healthcare delivery. Progress towards this goal has been made with the partnership of AMIA as an organization Member of CAHIIM. This interactive session is an opportunity for the audience to become involved in the value discussion, and to provide direct input into the CAHIIM Health Informatics Accreditation Council, Board and staff as it moves forward to update Standards, professional competencies, and procedures to meet the needs of an evolving and growing professional discipline.

**References**

1. Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM) website: [www.cahiim.org](http://www.cahiim.org)
2. Commission on Accreditation for Health Informatics and Information Management Education: Health Informatics 2010 Standards and Interpretations for Accreditation of Master’s Degree Programs in Health Informatics available at: [http://www.cahiim.org/applyaccred_HI_grad.html](http://www.cahiim.org/applyaccred_HI_grad.html)
Career Opportunities for the Many Paths to Informatics

Laura Wiley, MS1, Tiffany Kelley, PhD, MBA, RN2, Virginia Lorenzi, MS3, Vishnu Mohan, MD4, Jessica Tenenbaum, PhD5, Julie Doberne4
1Vanderbilt University, Nashville, TN, 2Nightingale Apps, Boston, MA, 3Columbia University, New York, NY, 4Oregon Health and Science University, Portland, OR, 5Duke University, Durham, NC

Abstract
Informatics is a diverse field with an even more diverse workforce. A demand for informatics professionals continues to increase, the number and types of informatics training opportunities also increase. From traditional informatics training programs to cross-training in informatics and informatics certification programs, there are a number of ways into an informatics career. Following our successful tradition, the AMIA Student Working Group proposes a “career panel” to offer perspectives and advice for students on career opportunities and professional development. This year, the panelists include a board certified physician informatician, an informatics PhD working in academia, a cross-trained PhD-level informatician entrepreneur and a certified informatics professional with industry and academic experience. They will share their career and educational experiences and discuss upcoming trends in informatics careers. This panel will help future and current informatics students and early-career professionals to better prepare for and develop their careers.

General Description
The Student Working Group (ST-WG) of the American Medical Informatics Association (AMIA) annually organizes a career oriented panel at the AMIA Fall Symposium to provide students with important perspectives and advice on career opportunities in informatics. In place since 2002, these panels have had various themes including the career stages of informaticians (2012), careers in academia vs industry (2013) and emerging informatics careers (2014). These panels are routinely well attended and serve as a highlight for student attendees.

Career opportunities for informaticians are growing exponentially and it is clear that there are a variety of training mechanisms available to individuals interested in the field. From traditional Masters and Doctorate informatics training programs to the new American Board of Preventive Medicine (ABPM) Clinical Informatics Subspecialty Fellowships, there are a variety of formal informatics training options. Additionally, due to the diversity of informatics a number of informaticians have training in another discipline with extensive informatics experience and expertise but without a formal informatics degree. There are also a number of informatics certifications from the aforementioned ABPM Subspecialty to Certified Associates/Professionals in Healthcare Information and Management Systems (CAHIMS/CPHIMS). This panel will feature speakers with diverse career and informatics training experiences to give both current and prospective students insight into career opportunities across a variety of types of informatics training.
This year we will build on our successful tradition and enhance it in the following ways:

(a) Feature panelists who have a variety of educational experiences in informatics including formal and informal informatics training
(b) Feature panelists who hold a variety of certifications including: CPHIMS and the recent ABPM Clinical Informatics Subspecialty.
(c) Cover a variety of career types from traditional tenure-track academics, to other academic and business/entrepreneurial roles.
(d) Provide a mixture of domain experiences from clinical informatics, bioinformatics and nursing informatics

In response to the large number of questions typically received at this event, we plan to take advantage of the new interactive panel format. Each of the panelists will have 5 minutes to introduce their educational and career background. The remainder of the session time will be allotted to audience and moderator driven discussion questions. Potential topics for discussion include:

(a) How has your educational training prepared you for your career?
(b) What influenced your decision to enter into the field of informatics and pursue your informatics training?
(c) How should students prepare and ready themselves for the coming career opportunities and trends?
(d) How should early career informatics graduates continue to improve their skills and be successful in their careers?

Given the different career experience and education background of the panelists, this panel will provide valuable perspectives for professional, PhD and master students, as well as those just entering the field and determining which educational path is best for their informatics career.

Organizers

AMIA Student Working Group Executive Committee and Volunteers
Laura Wiley, MS, Chair
Julie W. Doberne, Chair-Elect
Rui Zhang, PhD, Immediate Past Chair
Tiffany Kelley, PhD, MBA, RN, Student Representative to the AMIA Board
Katharine Fultz Hollis, MS, Member-at-large (Masters):
Onyinyechi U. Enyia Daniel, MA, Member-at-large (Doctoral-PhD)
Shauna M. Overgaard Member-at-large (Doctoral-PhD)
Andrew M. Harrison, Member-at-large (Doctoral-Professional)
Anthony Omosule, MS, Member-at-large (International)
Christopher Macintosh, RN, BSN, Student Representative to the Nursing Informatics Working Group
Panel Participants

Physician Informatician
Vishnu Mohan, MD, MBI, MBCS, FACP
Assistant Professor, Department of Medical Informatics and Clinical Epidemiology
Oregon Health and Science University, Portland, OR

Informatics PhD
Jessica Tenenbaum, PhD
Associate Director for Bioinformatics
Duke Translational Medicine Institute, Durham, NC

Nursing PhD
Tiffany Kelley, PhD, MBA, RN
Founder & CEO
Nightingale Apps, Boston, MA
Part-Time Faculty, Bouve College of Health Sciences, School of Nursing
Northeastern University, Boston, MA
Elected Student Representative to AMIA Board of Directors 2014-2016

Informatics Certified Professional
Virginia Lorenzi, MS
Associate, Department of Biomedical Informatics
Columbia University Medical Center
Manager, HIT Standards and Collaborations
NewYork-Presbyterian Hospital

Moderator
Laura Wiley, MS
Chair, AMIA Student Working Group
Vanderbilt Genetics Institute
Vanderbilt University, Nashville, TN

Statement
I, Laura Wiley, confirm that all panelists listed in this proposal have agreed to participate in this panel. Panelists are aware that there are no travel or registration funds available. Panelists are also aware that the Student Working Group is unable to reimburse their registration costs.
An Interactive User Interface for Drug Labeling to Improve Readability and Decision-Making

Hamed Abedtash¹, PharmD, Jon D. Duke, MD, MSc²

¹School of Informatics and Computing, Indiana University, Indianapolis, IN; ²Regenstrief Institute, Indianapolis, IN

Abstract

FDA-approved prescribing information (also known as product labeling or labels) contain critical safety information for health care professionals. Drug labels have often been criticized, however, for being overly complex, difficult to read, and rife with overwarning, leading to high cognitive load. In this project, we aimed to improve the usability of drug labels by increasing the ‘signal-to-noise ratio’ and providing meaningful information to care providers based on patient-specific comorbidities and concomitant medications. In the current paper, we describe the design process and resulting web application, known as myDrugLabel. Using the Structured Product Label documents as a base, we describe the process of label personalization, readability improvements, and integration of diverse evidence sources, including the medical literature from PubMed, pharmacovigilance reports from FDA adverse event reporting system (FAERS), and social media signals directly into the label.

Introduction

Food and Drug Administration (FDA) prescribing information (also known as product labeling or labels) are primarily meant to be used by physicians and other healthcare professionals to find accurate, reliable drug information needed to prescribe medicines safely and effectively. The labeling provides evidence-based information on adverse reactions, contraindications, dosing modifications in children and the elderly, and guidance regarding use in pregnancy and nursing¹. However, labels have also been extensively criticized for having too much information, making them complex, lengthy, and challenging to readers seeking relevant and important safety warnings²-⁴. In our previous quantitative analysis³, we found the “overwarning” problem with the adverse drug events (ADE) in the labeling content which hinders clinicians from grasping the significant safety alerts in the labels.

When reviewing a drug label, health-care providers must scan the entire label content to find relevant safety warnings, a cumbersome and time consuming task. While the introduction of label Highlights has brought some of this information to the fore, these facts are limited to the most severe or frequent events in the overall population, without consideration of patient-specific factors. Providers utilizing labels are most commonly seeking a particular piece of information related to a particular patient, whether based on co-morbidities, co-medications, or symptoms that have been raised in a visit. Finding such information with a label is inefficient and often impossible within the time constraints of a typical office visit. There are, of course, myriad websites and applications that provide drug safety information. However, the FDA-approved label is unique—it is thoroughly vetted and approved by a standardized process, and is recognized by manufacturers as the defining statement regarding their drugs. Thus, working with the actual label to improve its usability is a necessary and novel advancement in providing a better experience for providers while maintaining the ‘blessing’ of manufacturers and the FDA.

In this paper, we describe the development of an interactive, web-based application known as myDrugLabel. The goals of this project were to 1) improve the readability of FDA-approved drug labels by increasing the ‘signal-to-noise ratio’, 2) provide meaningful information to the physicians based on patient-specific comorbidities and concomitant medications, and 3) support the decision-making process by integrating additional evidence sources (e.g., medical literature, pharmacovigilance reporting, social media signals) directly into the label. In the following sections, we will provide an overview of the application design process and walk through the resulting application’s architecture and features through a sample clinical scenario.

Methods

Requirements and Design

We applied an iterative design process to prototype, develop, and refine the application components including data processing algorithms, user interface, and database. In each iteration, the developed functionalities were tested and necessary modifications were made. Healthcare professionals, e.g., physicians, pharmacists, and nurses were identified as target users of the application with a possible extension to patients as we add enhanced usability.
In the first step, the functional requirements of the application were elicited through a group composed of one physician and two pharmacists (including authors HA and JD) to define use cases and task scenarios. Figure 1 represents the use cases defined by this group, including label searching, modification of patient-specific factors, and integration of data from multiple sources. The requirements called for five data ‘providers’, the user and four external Application Program Interfaces (APIs), i.e. RxNORM, PubMed, FDA Adverse Event Reporting System (FAERS), and Topsy. The diagram represents the functional requirements for the system and the interactions between actors (e.g., user) and use cases (e.g., search a drug label). The “include relation” indicates that the included use cases are part of a larger use case and executed upon parent use case is called. The “use relation” or large hollow triangle arrow shows that the process of doing a use case (e.g., Customize Drug Label) always involves executing another use case (e.g., Add/Change Age).

Figure 1. Use case diagram of myDrugLabel application. The “include relation” specifies that the use cases are executed as part of a parent use case, and the “use relation” or large hollow triangle arrow indicates that execution of the base use case involves performing another use case. API=application program interface, FAERS=FDA adverse event reporting system, SPL=structured product labeling.

In the next phase, prototypes of the user interface were developed based on these requirements. Both designs had data entry sections on the left side; however, the design with a simpler interface in which supplementary information was displayed in tooltips above the highlighted terms was felt to be a better experience by the test users. The prototypes were evaluated by external users, and changes were made to the prototype interface based on the comments.

Technology Platforms

The proposed architecture was based on standard medical terminologies, open-source software and database platforms, and publicly available application programming interfaces (APIs). RxNORM® and MedDRA® (Medical Dictionary for Regulatory Activities) were used to map drug names and adverse events within the drug labels,
respectively. HTML5, CSS3, JavaScript, jQuery library, and AJAX were used to develop interactive webpages with asynchronous data transfer capability. Data were requested and transferred via PHP (server-side language for web application development) APIs from a MySQL database in extensible markup language (XML) or JavaScript Object Notation (JSON) formats.

Structured Product Labeling (SPL)

Structured Product Labeling (SPL) is the standard format accepted by FDA for transferring drug labeling contents between information systems. FDA drug labels were retrieved in SPL format from DailyMed and transformed to HTML files using the latest SPL schema available on FDA website, retrieved on October 14, 2013.

Processed SPL (P-SPL) Documents

In order to improve the performance of filtering label contents by diseases and comorbidities, a Python application was developed that identified adverse event terms in the SPL based on MedDRA hierarchy using simple natural language processing (NLP) and regular expression techniques to encapsulate in a span HTML tag. This approach saved time and decreased CPU load for the client facing interface. For example, the term 'myalgia' in the SPL was converted to `<span class="condition" mdrptid="36516837">myalgia</span>`. This modification allowed for rapid manipulation of the visual design of the label based on user input, without needing to search for terms in real-time (i.e., user waiting time). The response time highly depends on the client system configuration, length of the drug label, and how many times the term is mentioned in the label content. In average, the modification reduced the user waiting time from 60 to 2 seconds on a Desktop PC with CPU i5-3470 3.3 GHZ and RAM of 4MB.

Supplemental Evidence and Information

One of the recognized challenges to interpreting adverse events in drug labels is the lack of external evidence. For example, is an event listed simply because it was reported to the FDA, or have there been substantive studies exploring the associated risk between drug and event? We thus sought to supplement the labels with additional external evidence sources, displayed in a tooltip after selecting a drug adverse reaction or drug name. This was accomplished through the use of APIs pulling in data from the literature (PubMed API), the FDA adverse event reporting system (FAERS) (openFDA API), and Twitter posts (Topsy API). The information was streamed via respective APIs in a real-time manner.

FAERS data were retrieved from openFDA platform launched in June 2014. The queries to the API comprised two sections. In the case when an adverse reaction is selected on the label, the label’s drug name along with the MedDRA preferred term of the reaction are concatenated. Throughout our preliminary testing the API, we learned that both generic name and brand name of the drug should be in the query to fetch all records. For example:


Likewise, when a drug name on the label is selected, the chosen drug name and the drug label’s name are sent to API. The returned counts do not necessarily indicate the number of drug-drug interaction events. For example:


Results

Application Architecture

The architecture of myDrugLabel is a Service Oriented Architecture (SOA) design in which the client- and server-side components are connected via RESTful web services with the server database and external APIs (Figure 2). The client-side includes webpages, JavaScript codes, and P-SPL document. The server-side component is composed of a database, the bank of P-SPL documents, and PHP APIs.

The application has two pages, Home and Label View. The user starts looking up a drug name on Home page to select a drug of interest. Then, the page navigates to Label View page to display the labeling content and filtering functionality. In order to maximize speed of the client-side user interface, all data transfer processes between the server and browser are done in an asynchronous fashion. The only exception is when the drug labeling content is retrieved from the server. On the Home page, a request is sent asynchronously to the API once the user searches a
drug name, then the top 10 drug names similar to the input string are returned to fill the drop down list. Selecting the drug name directs user to the Label View page, which is the main page for processing the labeling content. The selected drug’s P-SPL document is loaded from server via the SPL Loader API and gets transformed to an HTML file by using the FDA-provided XSLT schema.

The application does not have a stand-alone business layer. Part of the calculations and text processing are performed on the client browser by JavaScript functions, including highlighting or hiding content sections, terms, displaying notifications on the right side, and tooltip functionality. Some other processes that needed security constraints, such as database connections and external API data transfer were designed to be performed by the server-side PHP APIs. The server APIs include SPL Loader (fetches SPLs from server by the requested SPL ID), Drug List Retriever (gets drugs lists from the database), MedDRA Checker (returns comorbidities’ synonyms from MedDRA dictionary located in the database), PubMed Retriever (requests and processes the literature statistics from PubMed), FAERS Retriever (requests and returns ADE statistics from FAERS), and Topsy Retriever (requests, calculates, and returned the number of signals on Twitter).

![Simplified view of myDrugLabel architecture](image)

**Figure 2.** Simplified view of myDrugLabel architecture. AJAX=asynchronous JavaScript and XML, API=application program interface, FAERS=FDA adverse event reporting system, MedDRA=medical dictionary for regulatory activities, PHP=server-side language for web application development, SPL=structured product labeling.

**Application User Interface**

Once a drug is selected from the Home page (Figure 3), the page is directed to Label View page (Figure 4), which is the primary location for interactivity in the application. On this page, the user can enter patient-specific parameters, which dynamically alter the labeling content (Figure 4, section A). For example, if the patient’s age is greater than 65 years old, the geriatrics section is automatically highlighted and the pediatrics sections is become hidden. Similarly, upon selection of male gender, the pregnancy and nursing sections are hidden. Making these adjustments enhances the readability of the label by reducing unnecessary content. Similarly, when a patient condition is added,
all synonyms of the disease are retrieved via the MedDRA API and highlighted in the label text (Figure 4, section B). Highlighting of conditions is limited to the following safety-related sections: Adverse reactions, boxed warnings, clinical studies, contraindications, warnings and precautions, general precautions, other safety information, pregnancy, risks, use in specific populations, and user safety warnings.

Figure 3. Home page of myDrugLabel application. User searches for a drug name in the provided textbox; once a drug is selected, it directs to Label View page.

To the immediate right of the label content is a customized navigation bar (Figure 4, section C) that provides quick access to highlighted sections of the label (based on patient parameters, such as pregnancy information or geriatric dosing instructions). The navigation bar also provides rapid access to highlighted terms within the label, such as relevant conditions, medications, and adverse effects. To help the user keep track of this content, related concepts are color coded throughout both the content and the navigation areas. Thus, for instance, if the user enters a condition of ‘Myalgia’, all information related to the condition will be highlighted in the same color throughout the document. Myalgia-specific highlights can then be toggled on and off as a group. For certain sections, color-coding is applied in standard fashion for all labels. For example, the pregnancy and nursing sections are highlighted in pink.

Integrated evidence from external data sources was shown via on-demand tooltips, for better accessibility and to avoid information overload. The tooltips are accessible by clicking on any highlighted term, either comorbidities or co-medications. The tooltip is composed of four tabs (Figure 5), including an evidence overview, literature data, pharmacovigilance reporting data, and social media signals. The overview tab gives the user a general sense of the strength of association between the selected term (comorbidity or co-medication) and the drug of interest. On the literature tab, the statistics of published papers mentioning both the drug (generic and brand) and the condition or its synonyms are presented stratified by type of publication, including journal article, case reports, clinical trials, controlled clinical trials, meta-analyses, observational studies, randomized controlled trials, and review articles. The link to PubMed search page is also provided, which will take the user to a new window containing the PubMed search results specific for that drug, condition, and article type.

The ADR reports tab displays how often the comorbidity or co-medication has been reported by other consumers of the drug of interest. These reports are divided into categories based on seriousness: 1) congenital anomaly, 2) disabling, 3) hospitalization-associated, 4) life-threatening, 5) death, 6) all serious events, and 7) all non-serious events. The social media tab provides real-time data on mentions of the drug and condition of interest (or its synonyms) on Twitter. The number of posts are grouped by 1-day, 7-day, 30-day, 6-month, and 1-year time frames.

Usage Scenario

An example scenario for myDrugLabel is demonstrated in Figure 4. A 71-year old woman is taking Lipitor for her hypercholesterolemia. She has also other comorbidities including diabetes, myalgia, hypertension, and heartburn for...
which she is taking acetaminophen, metformin, hydrochlorothiazide, and cimetidine. After the patient data is entered, the sections related to geriatrics turn to green, while pediatrics, pregnancy care, and nursing sections are hidden. The numbers next to the comorbidities and co-medications on the filter and notification bars show how many times they are mentioned in the label’s content. As shown in the figure, the sections that contain the terms are highlighted such as warnings, precautions, adverse events, and clinical trials, making it easy to find the relevant information specific to a particular co-morbidity or symptom.

![Image]

**Figure 4.** The Label View page. Using the filter box (A), user is able to customize the label content (B) by entering age, gender, pregnancy and nursing status, comorbidities and co-medications on the left side. On the right side (C), the user is notified of found sections and terms related to the input data. The navigation (D) bar helps user jump quickly to the point of interest within the content.

**Discussion and Future Directions**

In this project, we designed, prototyped, and developed an interactive user interface for FDA-approved drug labeling called myDrugLabel, reflecting the “personalized” aspect of the application. This web-based application promises improvement in readability of drug labeling and support for evidence-based decision-making. A formal evaluation study of the tool is underway. However, it has several novel features in its architecture and visual design. First, it utilizes the original FDA-approved product label as a base for all enhancements, injecting personalized highlighting.
and multiple external evidence sources into the original labeled content. As a result, the user can review the original
label, exactly as FDA approved, or by entering a few patient parameters, review a shorter, focused document
highlighting the specific areas of interest to the practitioner.

MyDrugLabel incorporates a number of existing APIs to pull real-time data from evidence sources, then further
stratifies this information to maximize usability. Most critically, myDrugLabel assists the provider in determining
whether a labeled side effect, potentially one of hundreds listed, is in fact supported by other sources. This
knowledge is critical to informing both the patient and prescribing decisions.

At the time of developing the website, the SPL information were accessible on the National Library of Medicine’s
Download Labels site (DailyMed). In August 2014, openFDA released drug product label API
(https://open.fda.gov/drug/label/) that provides access to over 25,000 prescription and 36,000 over-the-counter
(OTC) structured product labeling records. The new API has three main advantages. First, SPL information can be
retrieved with high level of granularity, meaning that the user has the option to download pieces of the label, e.g.,
Boxed Warning, Adverse Reactions, Drug Interactions, Patient Counseling Information, etc. Second, it is possible to
perform simple SPL mining using the API, for instance, to find the medications that have a specific ingredient or
adverse reaction. Third, the API has the power to perform a descriptive analysis, e.g., to identify the number of
products having a specific food interaction with alcohol or grapefruit juice. Considering these capabilities, the new
iterations of myDrugLabel will be a proper showcase of applying openFDA drug product label API.

A limitation of this paper is the lack of usability metrics to quantify the benefits experienced by users of this tool.
However, many of myDrugLabel’s features, such as simply removing pregnancy information from labeling for male
patients, shows the intuitive benefits of its design. Another limitation is that the presence of abstracts in the literature
or mentions on Twitter do not necessarily imply a causal or direct association between drug and effect. The search
strategies for these sources need to be refined to better interpret the evidence for users of the application.

Several next steps are planned for the development and evaluation of the application. First, as mentioned, we will
evaluate the application with healthcare professionals from different specialties as well as patient users. Secondly,
we will incorporate evidence from observational data networks to provide more clinically sound evidence of
associations between drug and adverse events. Finally, myDrugLabel will be released for public use, both to gather
feedback on larger scale and to incorporate patient-reported adverse effects into the evidence base.
Figure 5. The tooltip element provides supplementary information about conditions and co-medications. A) The Overview tab shows total number of items in each evidence section group. B) The Literature tab presents a stratified count of papers in the literature citing both the drug of interest and the selected term. C) The ADR Reports tab shows the number of reported cases associated with the drug and selected condition. D) The Social Media tab provides the count of Twitter posts mentioning the specific term (or its synonyms) and the drug of interest.
References

Content and Usability Evaluation of Patient Oriented Drug-Drug Interaction Websites

Terrence J. Adam, RPh, MD, PhD1,3, Joseph Vang2
1Institute for Health Informatics, University of Minnesota, Minneapolis, MN
2College of Liberal Arts, University of Minnesota, Minneapolis, MN
3College of Pharmacy, University of Minnesota, Minneapolis, MN

Abstract

Drug-Drug Interactions (DDI) are an important source of preventable adverse drug events and a common reason for hospitalization among patients on multiple drug therapy regimens. DDI information systems are important patient safety tools with the capacity to identify and warn health professionals of clinically significant DDI risk. While substantial research has been completed on DDI information systems in professional settings such as community, hospital, and independent pharmacies; there has been limited research on DDI systems offered through online websites directly for use by ambulatory patients. The focus of this project is to test patient oriented website capacity to correctly identify drug interactions among well established and clinically significant medication combinations and convey clinical risk data to patients. The patient education capability was assessed by evaluating website Information Capacity, Patient Usability and Readability. The study results indicate that the majority of websites identified which met the inclusion and exclusion criteria operated similarly, but vary in risk severity assessment and are not optimally patient oriented to effectively deliver risk information. The limited quality of information and complex medical term content complicate DDI risk data conveyance and the sites may not provide optimal information delivery to allow medication consumers to understand and manage their medication regimens.

Introduction

Adverse drug events, including injury and death, have been reported to affect up to 1.5 million patients annually, according to an Institute of Medicine report on Preventing Medication Errors. Exact numbers are difficult to assess, due to lack of uniform reporting, but adverse drug events are estimated to add 2-4 days to hospital length of stay at a cost of $2500-$5500 per patient1,2,3. A particular subset of adverse drug events, drug-drug interactions, occur frequently among community and ambulatory care settings with estimates on patient risk of exposures to drug-drug interactions ranging from 2.2% up to 70% of patients, depending on the population studied and the methods used in the study.4,5

Drug-Drug Interaction (DDI) information systems are essential tools to identify medication combinations which may place patients at risk of adverse drug events (ADE) including the potential risk of major morbidity and mortality6. DDI information systems have the capacity to help identify, analyze, and facilitate the avoidance of potentially dangerous or even lethal drug interactions both as stand-alone applications or as integrated decision support in electronic medical record systems. The availability of DDI systems to health professionals includes clinical databases, DDI assessment systems, web-based tools and clinical phone applications that evaluate patient risk via information delivery and clinical decision support7,8,9,10.

Clinical content on websites provide easy access points for patients or drug consumers to efficiently gather the necessary information to understand medication therapy concerns. However, if patients receive clinical information which lacks important clinical context they may be at risk of potentially unwanted adverse drug events due to the lack of information or inaccurate information. According to the Food and Drug Administration in the year 2012, more than 325,000 serious health outcomes and 54,000 deaths resulted from ADE, providing an important concern on the safe delivery of medications in healthcare11. With the estimated prevalence of DDI mediated ADE varying widely, it is not well established as to what percentage of ADE are directly associated with drug interactions, however, the DDI are a major clinical concern and patient safety opportunity since many of the DDI-mediated ADE are potentially preventable.

Patient oriented DDI Information systems are publically available computer databases available via websites, which have been created to inform patients and drug consumers about relevant drug interactions to help drug consumers...
stay informed on their medications and assist in ADE prevention. The typical clinical provider-oriented DDI
information systems provide information on potential DDI, drug-food interactions, or even drug-alcohol interactions
to help the clinician understand potential risk. The purpose of this study is to assess whether free and publically
available patient-oriented DDI assessment websites function in similar ways and contain accurate and reliable
information concerning DDI risk. Furthermore, the study will assess if these publically available websites contain
DDI risk content which is patient oriented in the risk communication information content (multimodal with graphics
and risk cues) and the required reading level for the estimated risk information (6-8th grade reading level).

Background

Current DDI information systems exist primarily in professional settings such as community and hospital
pharmacies and electronic prescribing systems for use in clinical decision support by medication prescribers and
dispensers. The presence of information systems has been successful in identifying potential DDI exposures at the
point of care. In an extended study done by Abarca et al., the study evaluated the performance of DDI screening
software in identifying clinically significant DDIs in pharmacy computer systems in community and hospital
pharmacies. They used 25 medications with a total of 37 drug pairs and 16 clinically significant drug pairs to create
6 mock patient profiles.12 Although there are improvements in the DDI screening systems, additional aspects needed
to be addressed included the ability to correctly detect clinically significant DDI in hospitals and understand the
exact severity of the potential interactions13. Later assessment of clinical pharmacy systems also show some
concerns still persist in DDI evaluation in the pharmacy setting.14

DDI information systems are common in pharmacies; however the data from these DDI information systems must
also interface with information systems from Pharmacy Benefit Management (PBM) companies and electronic
prescribing software15. In research work focused on the prescription claims data base of a PBM company, Malone et
al. performed a retrospective cross-sectional analysis of 46 million participants for 25 clinically potential significant
DDIs.16 The study covered a 25 month time frame and showed that an estimated 374,000 participants may have been
exposed to clinically important DDIs. The study also showed higher potential DDIs for individuals based on gender
and age differences. These results indicate that even with the presence of available information systems for DDI
risk screening, patients may still end up using at-risk medication combinations.

In assessing potential DDI risk, scoring systems have been developed to identify the severity of the potential DDI.
Prior work by Oshikoya et al. determined drug-drug interactions and severity ratings for patients with the human
immunodeficiency virus. They categorized drug-drug interactions and assigned a score on a scale of A (no known
interaction); B (minor/no action needed); C (moderate/monitor therapy); D (major/therapy modification); and X
(contraindicated/avoid combination).17 In addition, their results showed discrepancies between databases used such
as Medscape and USA MIMS. Having a means to score the potential DDI risk is needed for clinical decision
making to weigh the potential risks and benefits of medication combinations.

Previous efforts have indicated a lack of effective DDI software screening in the pharmacy setting and limited
ability to avoid clinically significant DDIs with PBM care delivery. Even with this knowledge, there has been a lack
of research focused on the DDI mediated adverse drug effects in the outpatient setting. The research approach is an
evaluation of publically available information for patient and/or caregivers concerning DDI. The goal of the study is
to analyze individual websites and determine if currently available information is valid and accurate with regards to
clinically significant DDI for drug consumers. The approach of the our research follows a similar approach to
Oshikoya17 in classifying variations that may exist between information content concerning severity ratings of
important DDI from the project websites. The significance of this research is to determine if the DDI information
offered to the general population has plausible clinical risk assessment guidance and provides the necessary
information to identify clinically significant DDIs. As previous research has highlighted, discrepancies in DDI
screening systems are present in both pharmacies as well as PBMs, this research work aims to assess for variations
in the information given to consumers about DDI risk.
**Methods**

**DDI websites.** The study focused on identifying available websites which had the capacity to inform medication consumers on potential DDI risk and were freely available to medication consumers. A total of 65 eligible websites were identified with expert selection and the search engine Google using search terms including drug-drug interaction, drug-drug interaction checker and drug-drug interaction assessment search terms. The search engine identified sites were supplemented with additional websites identified by expert informaticists and clinical practitioners. The majority of the sites were identified from the Google search effort. The identified sites were found and confirmed to be online and available during the July to September 2014 time period of primary data collection.

*Inclusion Criteria:* Websites were included in the study sample if they were freely available to external end-users with direct access to the site or with the use of an input email address for user authentication and potential use for profile tracking. English language sites were included in the study group.

*Exclusion Criteria:* The necessity for payment information or the lack of DDI checking functionality resulted in study exclusion. The websites must also have each of the clinically significant DDI medications available in their databases at the drug level. If the level of drug matching was only available at the therapeutic class level, the websites were excluded. All non-English language sites were excluded from the study. A total of 44 websites were included in the study after applying the inclusion and exclusion criteria.

**DDI medication selection.** The drug combinations in this descriptive, cross-sectional analysis of online DDI websites included a total of eight drugs and five DDI pairs which incorporated medications for multiple disease states (e.g. high cholesterol, fungal infection, etc.) to provide relatively broad medication coverage of key therapeutic areas at potential risk of DDI events (Table 1). The clinically significant DDI drug pairs were selected based on the previously published literature\(^1\) which was developed from expert consensus and large database validation. The selected medication combinations included high risk, high prevalence and potentially preventable drug combinations which could be expected to put patients at risk of adverse events. The presence of alerts for therapeutic class duplication was also measured through the Selegiline-Phenelzine drug pair to identify sites having the ability to assess multiple drug profiles where all the medications in the study set were built into a comprehensive drug profile. Some websites did not have multiple drug assessment (3 or more medication) capacity and could only assess binary medication pairs. Since all the drug pairs have been well established to be clinically significant DDI, it was expected that each should provide an alert when entered into the websites with similar estimated severity risk.

<table>
<thead>
<tr>
<th>Table 1: Drug Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simvastatin &amp; Itraconazole</td>
</tr>
<tr>
<td>Warfarin &amp; Gemfibrozil</td>
</tr>
<tr>
<td>Warfarin &amp; Levothyroxine</td>
</tr>
<tr>
<td>Fluoxetine (SSRI) &amp; Selegiline (MAO-I)</td>
</tr>
<tr>
<td>Selegiline &amp; Phenelzine (MAO-I)</td>
</tr>
</tbody>
</table>

For website data entry, the medications were recognized by the generic name as well as brand names and could be recognized as an oral tablet or oral solution in regards to the route of medication. All websites included in the study were able to correctly identify either the generic or brand name of each drug for the interaction assessment.

**Data Collection**

The DDI website information performance was assessed for clinical content and patient usability. The first method assessed the clinical content of DDI information system severity ratings offered to patients. Each website visited was noted for the presence or absence of a DDI severity rating. Four ratings were identified including contraindicated, major, moderate, or minor interactions. In cases where there was not a specific standard definition on the website, a normalized approximation of the standard score was identified from the DDI data and mapped to the four level standard. The patient oriented data assessment included information capacity scores, patient usability, and readability.
**DDI Information Capacity Score:** For the information capacity score, each website was quantitatively graded on the criteria including: presence or absence of a drug-drug interaction alert, presence or absence of severity grading, severity rating scales explanation, presence or absence of drug-food interaction, presence and absence of drug-alcohol interaction and therapeutic duplication. Each website that contained the complete information in regards to each of the five drug pairs for each of the categories received points. The websites were graded with five points in three categories; DDI presence, severity rating, and drug-food interaction. Another area graded, Drug-alcohol interaction, received a max of four points (one for each drug pair), however, because the drug pair simvastatin and itraconazole are recognized to have limited interactions with alcohol it was excluded from the scoring. The last category scoring was for therapeutic duplication which received a maximum score of two points. The first point identified whether the sight was able to identify the duplication in drug therapy class and a second point for identifying the drug classes which were responsible. This point scale was used to determine the information capacity of the website and had a maximum total of 21 possible points. A summary of the scoring rubric for the DDI Information Capacity score is noted in Table 2.

<table>
<thead>
<tr>
<th>DDI Identified</th>
<th>Severity Rating listed</th>
<th>Drug-Food Reaction</th>
<th>Drug-Alcohol Reaction</th>
<th>Therapeutic Duplication</th>
<th>Max Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>0-5</td>
<td>0-5</td>
<td>0-4</td>
<td>0-2</td>
<td>21</td>
</tr>
</tbody>
</table>

**Patient Usability:** The patient usability score focused on how a consumer would approach a DDI assessment website and how they would assess if the content was patient friendly and contained clinically useful features. Ideally these features were present and represented best practices for drug information management and patient risk communication. One point was received in categories for alert icons, color coding, and severity rating scale presence, medication pick list, and login/profile management for a total of 5 possible points. Alert icons are important as they serve as a more universal image of communication which could help patients understand risk even if language barriers or reading level of the patient was limited. Color coding can help to differentiate severity data, with typical use of the color red being associated with danger and green lights associated with low risk. Severity rating scale data was deemed important in determining how dangerous a DDI exposure is to the patient. Medication pick lists are crucial for the accuracy of navigating and identifying correct drug interactions for the drug consumer. Having login/profile management for multiple drugs allows the user to easily remember the drugs they’ve searched and avoid the inconvenience of having to re-enter a long list of medications. All of these features provided grading parameters for the patient usability scores. Half points were received if the website failed to provide these features for all five drug pairs but had the functionality of some drugs. For most websites, these features were all or none phenomena and would work or not work for all of the drug pairs.

<table>
<thead>
<tr>
<th>Alert Icons</th>
<th>Color Coding</th>
<th>Severity Rating Scale</th>
<th>Medication Pick List</th>
<th>Login/Profile Save for Multiple Drugs</th>
<th>Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>0-1</td>
<td>0-1</td>
<td>0-1</td>
<td>0-1</td>
<td>0-5</td>
</tr>
</tbody>
</table>

**Patient Readability:** Patient readability was determined through the Flesch-Kincaid grading model. This grading model was established by Rudolph Flesch and Peter Kincaid based on the understanding that information is more understandable with fewer words and shorter sentences. The content of information for the simvastatin and itraconazole interaction for each website was the only information content which was graded for readability. The content relevant to the drug consumer was assessed using the Microsoft 2010 Flesch Kincaid scoring assessment. Due to the potential for the commercial or generic drug names to affect this grading model, each of the drug names were replaced with the generic word “drug” instead of simvastatin and itraconazole to decrease the grade levels that would have been achieved with those words in the consumer content. The Flesch Reading Ease score was also assessed for each of the sites to provide additional measures on readability.
Results

Severity Grading

The severity ratings measured from the websites demonstrated substantial variation. Approximately half of the websites did not provide severity rating for any of the drug categories. Although a substantial sub-population of the websites failed to give severity ratings, some indicated severity levels for the five drug pairs. For sites with severity ratings, about 50% indicated a moderate severity for the drug pair Warfarin-Gemfibrozil and the other 50% indicated a major severity for the same drug pair. For the Warfarin-Levothyroxine drug pair, 14 of the 44 drug pairs indicated a moderate risk associated with these two drugs while 5 of the 44 websites indicated that this drug pair had a major risk. For the Fluoxetine-Selegiline drug combinations, 12 websites noted that the interaction was contraindicated while five websites noted that it was a major severity reaction and one website noted that it was a moderate reaction. For the Selegiline-Phenelzine drug combination, nine websites noted that this drug pair was contraindicated; three websites noted that it was a major reaction, and one website noted that it was a moderate reaction. The DDI severity data is summarized in Figure 1.

Figure 1: Drug-Drug Interaction Severity Frequency

Information Capacity Scoring

Among the 44 websites included in this research, the mean Information Capacity score was 13.36 points with a standard deviation of 3.09 points. The highest scoring website had a score of 21, which was the maximum possible score. Two websites scored 19, which was just below the maximum possible scores. The lowest scoring website had a score of 4 and the second lowest score belonged to three other websites with a total score of 8 each. 42 of the 44 websites lacked the Therapeutic Duplication feature. The results for all the DDI Information Capacity scores are noted in Figure 2.
**Patient Usability Scoring**

The average score for the Patient Usability score for all 44 websites was 2.9 points with a standard deviation of 1.06 points. The distribution of the scores is noted in Figure 3. Only five websites received perfect scores for patient usability scoring. The scores that received half a point for some information features failed to have that specific feature for all drug pairs. The most common features for half points given were alert icons and severity rating scales.

Figure 3: Patient Usability Scores
**Website Readability Scoring**

The average Flesch-Kincaid grade level scored was 11.2 with the highest grade level being 16.2, the lowest grade level being 6.7 and a standard deviation of 2.29 with the results summarized in Figure 4. The Flesch Reading Ease score was also measured for the 44 websites was 40.27 with the highest score being 70.9, the lowest score being zero and the standard deviation being 15.11 with the results summarized in Figure 5.

![Figure 4: Reading Level of DDI websites](image)

**Grade Level Required To Understand Website Content**

![Figure 5: Website Reading Ease Measures](image)

**Flesch Reading Ease Scores**
Information on Drug-Drug Pairs Identified

Of the five drug pairs included in this study, three drug-drug interactions were not correctly identified by all of the DDI checkers. The two drug pairs, Simvastatin-Itraconazole and fluoxetine-selegiline, were correctly identified for a potential interaction by all of the DDI checkers. All but one website correctly identified the Warfarin-Gemfibrozil drug interaction. Another website failed to identify the Warfarin-Levothyroxine drug pair, but correctly identified all other drug interactions. 20 websites had identical results likely indicating a common DDI assessment engine in use.

Discussion

The study findings indicate that the many of DDI Checkers operate similarly or even identically as would be expected for a series of well-established DDI medications. Of the 44 websites, 20 had identical scores with an Information Capacity score of 13, Patient Usability Score of 2, and Flesch-Kincaid Grade Level of 12.4. The results also show that the majority of these websites may not be optimally patient-oriented with a low average Information Capacity Score and an average Flesch-Kincaid grading level of 11.2. The optimal Flesch-Kincaid reading score for patient oriented content would be lower with grade levels of 6-8 to be easily readable by most patients from a variety of educational backgrounds. The average Flesch Reading Ease was also lower than expected for ease of use by most readers and indicated that the information given to patients may be difficult to interpret.

The research provides important insights on publically available DDI information systems for well-established DDI combinations. The average Information Capacity Score of 13.36 implies that there are many features that are overlooked by the DDI checkers. Scores not received due to lack of a severity rating was the primary reason for the low scoring in Information Capacity. Other features that contributed to a decreased Information Capacity scores included the absence of the Selegiline-Phenelzine drug interaction, lack of presence for drug to food interactions, absence of drug-alcohol interactions, and absence of the therapeutic duplication data. Reasons for the lack of drug-food interaction and drug-alcohol interaction may have been due the fact that the websites focused on strictly drug-drug interactions or may have lacked the content in their database.

Another broader concern was the variability in severity ratings among the websites which provided severity data. The interactions used in the assessment are well known and clinically significant. Although the vast majority of sites indicated the presence of an interaction, the lack of severity data and lack of consistent severity data is potentially problematic. If a patient were to put in a medication combination which has a low risk of adverse clinical effects and gets no cues on severity and at the same time inputs one or more of the combinations included in this study, they may get the message that the low-risk combination is the same risk as the high risk combination. This has the potential to provide inappropriate messaging depending on how the patient perceives and internalizes the lack of risk severity. It is also easy to see a scenario where a patient may have a complicated regimen with multiple interactions and asks their physician or pharmacist about the interactions and gets reassurance on the first couple of interactions and then fails to ask or appreciate the risk for a high risk combination further down the list. Among the medications used in the study, there was also an unexpectedly high variation in severity rankings when it was provided. It would be expected that the severity data would be more homogenous, but in figure one the range went the full range of minor to contraindicated which is concerning from a patient safety standpoint and none of the pairs had full agreement on severity when it was provided. In addition, some of the interactions were not noted as DDIs on some sites.

The patient usability and reading levels could be addressed by insuring the content has multiple modes of risk communication with percentage data, risk symbols and use of color coding or other pictographic elements to convey potential risk when present. For the readability and reading ease, the DDI information system content could use more simplified explanations so as not to overwhelm the drug consumer with the information content and have less use of medical jargon. The high grade level to the content may have been due to the reuse of risk information developed primarily for use by health professionals.

Limitations to the study include a lack of prior studies in DDI information capability to use for scoring and assessment. Without a well-recognized grading criterion for patient oriented DDI information systems, generating one would require a consensus among professionals as to how important each grading factor was in evaluating these information systems. Future work to validate the scoring approach is being planned to assess scoring validity. Another limitation included a small sample of drug pairs as well as limited number of drug classes which were
studied which may have affected the study outcomes. There are many more clinically significant drug interactions that were not included in this study and further research could provide additional insight with an evaluation of those drug combinations. Ideally more drug combinations would have complemented the research, but with the time and effort needed to complete the thousands of potential drug interactions it would require a large labor effort. Furthermore, additional DDI combinations are likely to have diminishing returns. Also of note, many of the websites generated identical results and it would be expected that the pattern would continue with testing of additional drug pairs.

An additional limitation was the lack of actual patient feedback on the use of the systems. However, without systematically screening and assessing the quality of the information provided, putting patients with their actual drug regimens was deemed potentially problematic since counter-detailing may have been required based on clinical circumstances and the patient guidance provided by the sites. Assessing patient initiated mock drug profiles may be an alternative approach to get pertinent patient data on website usability.

Conclusion

The results of the research indicate that the content of these websites is similar but a large number of sites lack severity data, advanced drug profile management, and therapeutic class assessment. In addition, there is a broad problem of limited patient readability and limited patient oriented risk communication features. Based on the study results, further research and website development work is needed to improve patient oriented DDI information systems.

Possible approaches to improve patient oriented DDI checking sites could be the incorporation of a standardized severity grading scale to avoid potential confusion for the drug consumers on DDI risk. Improvement of patient oriented DDI information systems should also focus on simplifying drug interaction information provision to consumers with patient oriented content and inclusion of advance drug information management features.

References


OpenHealth Platform for Interactive Contextualization of Population Health Open Data

Jonas S Almeida*, Janos Hajagos¹, Ivan Crnosija¹, Tahsin Kurc¹, Mary Saltz², Joel Saltz¹
¹) Dept Biomedical Informatics, 2) Dept Radiology, Stony Brook University, State University of New York.
* jonas.almeida@stonybrookmedicine.edu

Abstract

The financial incentives for data science applications leading to improved health outcomes, such as DSRIP (bit.ly/dsrip), are well-aligned with the broad adoption of Open Data by State and Federal agencies. This creates entirely novel opportunities for analytical applications that make exclusive use of the pervasive Web Computing platform. The framework described here explores this new avenue to contextualize Health data in a manner that relies exclusively on the native JavaScript interpreter and data processing resources of the ubiquitous Web Browser. The OpenHealth platform is made publicly available, and is publicly hosted with version control and open source, at https://github.com/mathbiol/openHealth. The different data/analytics workflow architectures explored are accompanied with live applications ranging from DSRIP, such as Hospital Inpatient Prevention Quality Indicators at http://bit.ly/pqiSuffolk, to The Cancer Genome Atlas (TCGA) as illustrated by http://bit.ly/tcgascopeGBM.

Introduction

Technological and policy trends are reshaping the development and deployment of Biomedical Informatics applications. The increasingly data-intensive, patient-facing nature of modern Medical Systems are translating well established architectural recommendations, such as the decoupling between the data layer and presentation layer, into specific requirements. Furthermore, the ongoing evolution of the Web into a global data space (Heath 2011) is now a ubiquitous Big Data reality, with the Web Platform (Web Platform 2013) serving as its unifying computational environment.

As a matter of policy, public data held by government agencies in the US and many other countries must be available as Open Data (White House 2013). The technical requirements for Open Data play a foundational role in Web Technologies, defining a degree of openness along a 5 point system (Berners-Lee 2011), which ranges from being on the web (1 point) to using a linked Resource Description Framework (RDF, 5 points). Most public health data systems maintain a score of at least 3.. This is typically achieved by relying on a nonproprietary format, such as JSON, exposed through an HTTP REST API such as that provided by Socrata’s SODA data services (Socrata 2014). The availability of a wealth of health data resources satisfying this level of interoperability can be readily verified by pointing a web browser to State and Federal data resources such as health.data.ny.gov, data.medicare.gov, data.healthcare.gov, or data.cms.gov.

This report explores the feasibility of OpenHealth analytical applications that are entirely free from server-side presentation resources. The motivation for this approach derives from the
unparalleled scalability promised by “beyond the data deluge” analytical solutions based on code distribution (Bell 2009). This pursuit is also informed by a number of previous results showing that client-side code distribution is no longer bound by significant algorithmic or performance limitations. Regarding the latter, browser hosted computing was shown to scale well into the realms of HPC (Wilkinson 2014). Similarly, even specialized operations, such as image (Almeida 2012a) and sequence analysis (Almeida 2012b), have, for some time, been efficiently supported by the computational engines of the Web Platform. Nevertheless, it was also abundantly clear from the onset of the study that new capabilities were needed for functionalities conventionally hosted server-side such as data caching and normalization. A third challenge to the OpenHealth route was devising an approach to the user interface assembly that would allow researchers to present prototype analytical applications to domain users with minimal overhead.

Methods

Computational application - The OpenHealth platform was developed entirely in client-side JavaScript, which is made publicly available, and is publicly hosted, with version control and open source, at https://github.com/mathbiol/openHealth. The versioned hosting feature is achieved by versioned code development in GitHub’s gh-pages branch. The architecture of the OpenHealth platform is detailed in the Results section, and includes reliance on the native NoSQL data management and storage resources of the Web Browser - IndexedDB (http://www.w3.org/TR/IndexedDB). Code development, including the use of IndexedDB, was pursed with strict adherence to W3C standards recommendations. In principle, this should render OpenHealth cross-browser, but this behavior was only tested for Google’s Chrome and Mozilla’s Firefox browsers, albeit on multiple platforms: desktop, tablet and cell phones running multiple operating systems: Mac OS, Windows, Android, and Chrome-OS.

External data and libraries - The illustrative applications make use of specialized libraries for visualization, d3.js (https://github.com/mbostock/d3), and dimensional charting (https://github.com/dc-js/dc.js). Data caching was mediated through Mozilla’s open source localForage.js library (https://github.com/mozilla/localForage). These libraries were chosen, in part, because of OpenHealth’s emphasis on public versioned hosting of open source code. Similarly, all data used in the applications described in this report are available in the public domain from a variety of State and Federal agencies, primarily hosted by health.data.ny.gov and data.medicare.gov for case study examples 1-3. In all three cases, data retrieval is performed by demand of the analysis, using the Socrata JSON formatted open data services API (http://www.socrata.com/products/open-data-api) of those Open Data resources. The data normalization case study 4, which uses data from The Cancer Genome Atlas (Figure 5) illustrates a solution for patient-derived data resources made public without the complement of a data service API, or even of a Cross-Origin Resource Sharing (CORS) enabled web serving. The solution found, described in Results, is the basis for a number of recommendations in the Discussion. The TCGA data used is nevertheless in the public domain (no restricted access TCGA data sets were used), hosted by The National Cancer Institute of The National Institutes

Results
The main goal of this project was to develop a distributable platform for data-intensive computation of OpenHealth Data entirely as a client side Web Application. This design overcomes the need for server-side components and maximizes scalability through code distribution (Bell 2009). Open Government mandates and NIH dissemination requirements have engendered a wealth of reliable, high availability, Big Data web services. These web services and their patient-resolved clinical and biomolecular data are central to the interactive systems described, which illustrate the use of the OpenHealth platform (OH).

Architecture

Figure 1 - Simplification of the conventional server-side data management and normalization architecture (A) by relocating that functionality (B) to the OpenHealth Platform (OH) assembled within the browser (dynamically loaded as JavaScript libraries), using native, W3C standardized, data management resources such as NoSQL IndexedDB (http://www.w3.org/TR/IndexedDB). It is critical to note that in (B), all data and computation is performed by either the data providers or by the domain consumers, not by dedicated Biomedical Informatics computation infrastructure - which is no longer needed. In other words, all that the Biomedical Informatics application layer does is provide the JS code, distributed directly to the Web Browser hosted OH component, by the standard script tag injection mechanism.

Illustrative applications with <OpenHealth>?<analysis> URL composition
The OpenHealth platform’s data management and normalization functionalities are best visualized through illustrative examples. These illustrations will address four boundary scenarios: 1) Graphic interaction with a large data resource containing population-level data; 2) traversal of a large collection of individual patient data; 3) Cross-tabulation of multiple sources;
and 4) normalization of patient-derived biomedical data available in the public domain as raw data files. In each of these examples, the same code migration URL composition pattern is invoked. OpenHealth’s core library is dereferenced with the analytical code pulled in as an additional search parameter. For example, the interactive representation in Figure 2 can be produced from http://mathbiol.github.io/openHealth/?jobs/pqiSuffolk.js.

1) Graphic interaction with a large data resource containing population level data

Figure 2 - Pure client-side assembly of an interactive tool for Hospital Inpatient Prevention Quality Indicators (PQI) for Suffolk County, Long Island, NY: bit.ly/pqiSuffolk. The analysis begins with OH retrieving the data for each of the relevant 107 zip codes (out of 111474 described in the reference health.data.ny.gov/resource/5q8c-d6xq source for the state of NY), and caching them in the native NoSQL browser resource, IndexedDB (see Methods). Like other native data resources, such as localStorage and WebSQL, IndexedDB will persist between sessions for the domain name. As a consequence, subsequent accesses to this interactive Web Application will not exhibit the same initial waiting period for data retrieval.

To illustrate the point that <openHealth>??data analysis</openHealth> is just a generative pattern, and that the analysis code can be dereferenced from both relative or absolute URLs, it is useful to note
that the same result captured by the snapshot in figure 2 could be produced by http://mathbiol.github.io/openHealth/?https://rawgithub.com/SBU-BMI/openHealth/33a1b22d0e0f7fd786bfe2ebbf024ac55523262/jobs/pqiSuffolk.js. This formulation highlights the criticality of the versioned hosting feature of OH; the analysis URL points to a specific version, hosted in a different domain. That is, for the preventable disease interactive display coded by version UID 33a1b22d0e0f7fd786bfe2ebbf024ac55523262. The URL composition is convenient to illustrate this report, but a more useful approach may seek to engage the code migration functionality programmatically, as discussed later in this report.

2) traversal of a large collection of individual patient data

Figure 3 – Tabulation tool dereferenceable by shortcut bit.ly/sparcs2012. This example illustrates the logistics of traversing over 2.5 million de-identified hospital inpatient discharges in the state of NY in 2012 (Open Data resource health.data.ny.gov/resource/u4ud-w55t). The command line inset is a snapshot of the browser native tools, showing the values of 36 parameters for one of the 168,044 records found for Suffolk county. Note that first use of this interactive application will require pre-processing times of a few minutes, depending on the machine, but subsequent uses have OH shorten data retrieval to under half a minute, even in a moderately resourced mobile devices.
3) Cross-tabulation of care providers

Figure 4 - Example of an interactive application that crosses two Medicare databases for an arbitrary provider identifier and the corresponding hospital affiliations. Loading times should be nearly instantaneous on any device.

4) Normalization of patient-derived biomedical data made public as raw text files. The example application in Figure 5 tests the limits of what can be accomplished using the proposed approach when the patient-derived data is exposed without any specific provisions for web-based processing (see Discussion). This application retrieves two raw text files from https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_fupusers/anonymous/tumor/gbm/bcr/biotab/clin/, one describing the clinical data for 592 TCGA patients diagnosed with Glioblastoma Multiforme, nationwidechildrens.org_clinical_patient_gbm.txt, the other describing 1,284 pathology slide images obtained from the corresponding tumor samples, nationwidechildrens.org_biospecimen_slide_gbm.txt. While it may not lead to a noticeable delay in data retrieval and normalization, the inspection of the application code, migrated from github.com/mathbiol/openHealth/blob/gh-pages/jobs/tcgascopeGBM.js, will reveal the mediation of a cloud hosted proxy application, bit.ly/getTCGAxt, that pass the text content across TCGA data hosting domain restrictions (see Discussion).

Discussion

The interactive applications and supporting data processing configuration of the OpenHealth platform is proposed to be a good fit for patient-centric, outcomes-driven health delivery programs such as Medicaid’s Delivery System Reform Incentive Payment (DSRIP, bit.ly/dsrip) and the new Patient-Centered Outcomes Research Institute (PCORI). However, the domain
facing nature of modern Information and Communication Technologies (Almeida 2014) is, of course, not a reaction to these systems—quite the opposite. Healthcare is a latecomer to both the commoditized consumer-facing ICT (Mandl 2012) and the operational improvements facilitated using Big Data (Murdoch 2013, Manyika 2011). Therefore, it is reasonable to anticipate that the growing wealth of patient-resolved health data resources delivered by Open Government mandates, as articulated by the US Department of Health and Human Services (HHS 2015), will fundamentally change not only Health Information Systems, but the research and understanding of disease (Roth 2015). What data services and application development architectures will prove more effective in meeting those goals is the broader question that the OpenHealth platform described here explores.

Figure 5 - Snapshot of interactive application assembled from TCGA raw text files describing TCGA patients diagnosed with Glioblastoma Multiforme, and the pathology slide images obtained from the corresponding biopsies (see text for details). The cartesian plot on the right projects, interactively, the position of each patient by age and survival (days to death), while tracking the Karnofsky score (color) and images (diameter). The effect of histopathology and demographic features can be assessed visually through interaction with the corresponding bar charts.

The interactive applications in figures 2-4 provide evidence that OH’s server-less architecture, described in Figure 1B, is computationally more efficient, it is far easier to disseminate, distributing analysis that are therefore easier to reproduce. Although the applications in the Results section were assembled by a URL composition pattern, <openHealth>?<data
analysis>, the core OH library can be loaded programatically by script tag loading as in `<script src="https://mathbiol.github.io/openHealth/openHealth.js"></script>` or, for example, using jQuery, `$.getScript("https://mathbiol.github.io/openHealth/openHealth.js").` This is explained in detail in the project’s code development page at https://github.com/mathbiol/openHealth.

The interactive TCGA Glioblastoma (GBM) application described in figure 5, however, explores a more convoluted scenario, wherein the data is served to the public domain, but no API is available and cross-domain calling (CORS) is disabled. These obstacles have been noted by several reports over a number of years, but persist on many open repositories of public data. To overcome those barriers to interoperability with interactive web applications, we were compelled to develop a minimal proxy (server-side) mediator. To avoid a relapse to the troubled dependency on server-side resources (Fig 1A), we developed that TCGA proxy component as a Google Cloud hosted service, which can be inspected at bit.ly/getTCGAtxt. The TCGA GBM application was implemented to augment a web-based virtual microscope deployment, which hosts TCGA microscopy image data and image segmentation results. The application allows users to visualize various clinical attributes and select a subset of cases based on these attributes. Users can then interactively view these select cases using the virtual microscope platform. The need for the proxy component could have been entirely removed if the TCGA data web server included an open domain header (CORS), as noted elsewhere, namely (Robbins 2013) section 4.5 “Summary of technical recommendations for biomedical big data hosting”. On a more positive note, the normalization of the TCGA file contents did not itself present a major obstacle, as can be verified by noting the short loading/parsing times of bit.ly/tcgascpeGBM (Fig 5), and by following the cBio links (Cerami 2012) to those Web Applications.

**Conclusion**

OpenHealth is an in-browser JavaScript platform developed to mediate the management and normalization of Open Data in the Health Sciences domain in a manner that is scalable, secure and reproducible. Data pre-processing functionalities are conventionally associated with dedicated server-side resources. OpenHealth redirects that support to the native data resources of the modern Web Browser, which is now equipped with a computationally efficient JS interpreter, secured within a sandbox that isolates the execution of code migration from unauthorized access to additional local resources. This approach was found to be particularly effective for the development of interactive applications, and for the dissemination of reproducible analytical procedures. Mounting adoption of Open Government and Open Data mandates in Health Care suggests a key role for this architecture in measuring health outcomes and personalizing care delivery.

**Acknowledgements**

This work was supported in part by 1U24CA180924-01A1 from the NCI, R01LM011119-01 and R01LM009239 from the NLM. The authors also thankfully acknowledge support from Suffolk Care Collaborative Delivery System Reform Incentive Payment Program (dsrip.uhmc.sunysb.edu).
References

**An Associative Memory Model for Integration of Fragmented Research Data and Identification of Treatment Correlations in Breast Cancer Care**

Ashis Gopal Banerjee, PhD¹, Mridul Khan², John Higgins, MS³, Annarita Giani, PhD¹, Amar K. Das, MD, PhD³

¹General Electric Global Research, Niskayuna, NY; ²Department of Computer Science and ³Department of Biomedical Data Science, Dartmouth College, Hanover, NH

**Abstract**

A major challenge in advancing scientific discoveries using data-driven clinical research is the fragmentation of relevant data among multiple information systems. This fragmentation requires significant data-engineering work before correlations can be found among data attributes in multiple systems. In this paper, we focus on integrating information on breast cancer care, and present a novel computational approach to identify correlations between administered drugs captured in an electronic medical records and biological factors obtained from a tumor registry through rapid data aggregation and analysis. We use an associative memory (AM) model to encode all existing associations among the data attributes from both systems in a high-dimensional vector space. The AM model stores highly associated data items in neighboring memory locations to enable efficient querying operations. The results of applying AM to a set of integrated data on tumor markers and drug administrations discovered anomalies between clinical recommendations and derived associations.

**Keywords**

Associative memory, breast cancer treatment, electronic medical record, tumor registry, data integration, correlation

**Introduction**

While critical data on tumor pathology, genomic biomarkers, patient demographics, and clinical treatments are often gathered in different resources, investigators need all of these results to be brought together for comparative effectiveness, population research and translational science. For example, data linkage between the SEER cancer registry and the Medicare claims database has been valuable in discovering population-based patterns in cancer screening and treatment outcomes that would not be possible otherwise⁴. As part of the Oncoshare project⁵, a collaborative multi-institutional study of patterns, predictors, and outcomes of breast cancer care, data integration between local and state cancer registries with electronic health records revealed the limitations of relying on any single source for research. Before linking data among different systems into Oncoshare, the separate data sources revealed varying rates of cancer-specific diagnostic tests and treatments. For example, the rate of mastectomy in one hospital’s registry was 41% but the facility’s billing record only indicated that about 22% of patients received such surgical intervention⁶. After manual verification of the missing data, the rate of mastectomy across data sources was found to be 43%. In addition, linking revealed data patterns that were implausible, such as patients whose treatment dates from the EMR occurred a year before receiving cancer diagnosis in the registry.

Many academic medical centers are undertaking this data integration effort by creating data warehouses that incorporate research-related data from a variety of systems. Such an approach requires considerable planning and programming efforts that may require a year or more before the data warehouse is usable for research⁷. While data warehousing permits robust and rapid querying over uniformly represented data, this solution has a number of drawbacks for researchers as end users. Data warehouses typically do not incorporate all of the data in source systems, do not allow for multiple relationships to be maintained among source data, do not manage provenance of the data, and do not monitor changes in the schema or data quality in source data.

Several informatics groups have created new methods to tackle various aspects of this problem. For example, the Electronic Medical Records and Genomics (eMERGE) Network method aims to link samples and results collected from genetic studies to electronic medical records data to allow for high throughput biomarker studies⁸. The widely used Informatics for Integrating Biology and the Bedside (i2b2) data repository method allows for the merger of multiple sources of genetic, phenotypic, and other types of data into a single integrated schema that supports queries across data sources⁹. This prior work on data linkage and data warehousing methods provides data integration at the data or schema level, but not at the systems level. That is, these methods do not model or automatically adapt to
changes in the information systems that provide data. Consumers of integrated data may, as a result, not be aware of which systems provide data and when new systems have emerged or existing ones are altered.

In this paper, we explore the use of a novel data modeling approach, called *associative memory* (AM), which permits the rapid integration and correlation of data from multiple data sources\(^7\). Our analysis technique comprises an AM model\(^7\) that allows for a form of cognitive computing to mimic the human capabilities of processing, encoding, consolidating, and retrieving information from a constant influx of data streams captured via the sensory organs. We choose this model as it has the potential to address the volume, velocity, veracity, and variety (the four V’s of big data) challenges of data source agnostic aggregation and analysis. We apply the AM approach to rapidly integrate patient data collected from tumor registry and electronic health records, and conveniently identify the correlations between tumor markers, patient factors, and selection of administered drugs. While some of the identified correlations follow expected clinical patterns, others point to anomalies between clinical guidelines and drug administrations, indicating the need for further studies on the quality of drug administration recording as well as patient and physician adherence to the recommended guidelines. We also compare the simplicity of this approach against the more traditional effort needed with generating reports using SQL.

**Methods**

**Design and experimental setting:** Our breast cancer data sets come from two sources at Dartmouth–Hitchcock Medical Center (DHMC). The first source is a patient tumor registry and the second source is an electronic medical record (EMR) database, which is from the Epic vendor that was installed at DHMC on April 2, 2011. The current tumor registry has existed in its current form for the past decade. To evaluate our proposed AM approach, we obtained IRB approval to extract patient data from both the systems on November 1, 2014. The EMR data consisted of encounter data that recorded diagnostic and treatment information, including medication administration records. The tumor registry data consisted of stage (pathological, clinical, and combined) and grade information, along with tumor markers, specifically progesterone receptor (PR), estrogen receptor (ER), and HER2/neu.

Using the diagnostic date stored in the institutional tumor registry, we identified female breast cancer patients who had invasive breast cancer (Stage I-IV) after April 2, 2011, and we cross-linked these cases to their data within the EMR. Since we required treatment data and treatment is normally completed within a year, we limited the date of diagnosis to be one year before November 1, 2014. The data sets consist of 928 patients and 50,490 encounter records collected over the specified period. For automated de-identification of patient data, we identify each patient solely by a digital fingerprint that is created by a one-hash algorithm\(^4\). Using the digital fingerprint, we define a fixed, random temporal offset between plus or minus 30 days that is added to all the time stamps for a patient, effectively altering the true time of occurrence but maintaining the relative temporal distance between the events.

![Figure 1: Conceptual illustration of associative memory system showing data aggregation and analysis techniques](image-url)
While various AM models have been proposed, we adapt a specific implementation described in a 2011 patent that is efficient, easy-to-use, intuitive, and scalable. By replacing the data items that are conventionally stored in tables or tuples with elementary “atoms” of information residing in a common n-dimensional (n ~ 1 billion) vector space of contextual associations among the data attributes, this AM model (Figure 1) provides a unified and compact representation for any data type, dynamic or stationary, structured or unstructured, of arbitrary size and granularity.

The atomic pieces of information are represented as byte arrays of arbitrary sizes where the maximum size limit is enforced by the operating system constraints. The associations among the data items are naturally formed based on all the attributes such as the names, counts, hierarchical relationships, and qualitative and quantitative properties of the items. The quantitative properties are categorical strings, discrete integers, or continuous-valued floating point numbers, names and qualitative properties are represented as strings, and hierarchical relations are denoted as binary integers linking pairwise items. Each attribute type, other than hierarchical relations, then forms a dimension of the associative vector space (AVS) in which all the data items are organized.

Naturally, the data attributes vary widely among items representing fundamentally different entities such as chemotherapy protocols and surgical procedures, but are identical, albeit with different values, for the same entity. Various instances of identical entities lie in the same sub-spaces of the AVS, whereas instances of different entities occupy different sub-spaces of the common AVS. The sub-spaces for the different entities may overlap, indicating the presence of common attributes among them. Multiple occurrences of the same entity instance are represented as the same atomic piece of information with the provision to add more attributes, and, thereby, increase the dimensionality of the occupying sub-space. All the instances of a particular entity or similar entities are strongly connected to each other, thereby forming a natural cluster using a simple heuristic k-means method. This method

Figure 2: Simple user interface to define and run queries on breast cancer patient associative memory model
employs a hybrid Euclidean (for real-valued attributes) - Hamming (for string attributes) distance function as the similarity metric or the connection weights between the data items. Instances belonging to different clusters may also have some connections, but those are much sparser with lower weights. These connections are bi-directional and dynamic, thereby enabling the additions of new associations as more data is ingested.

While commonly contextualized (clustered) data entities are co-located in the organizational space, a virtual pointer-like token, called the relationship construct, provides the means to connect anything in the AVS to anything else. Each token uniquely identifies a particular atom of information, is the virtual location of the item in the AVS, and is the logical address of where the item exists on the physical storage medium. This capability ameliorates the requirement for physical co-location to articulate sub-spaces and instead uses associative nearness (shortest connecting path length) or dimensional proximity (number of overlapping sub-space dimensions) to enable endless clustering possibilities of data entities in an unlimited number of sub-spaces. This multi-faceted holographic-like framework allows for viewing of data from virtually any perspective without the need for additional processing. Furthermore, one no longer needs to search for identifying any form of associations among the data items as all such associations are maintained as tokens, co-incident with the related data items. Thus, the associations are obtained by merely referencing the items of interest and using the co-incident tokens to directly index the referenced items in the storage medium. This novelty allows for real-time correlation analysis as long as the data attributes are defined by the user. Figure 2 shows the user interface for interacting with and querying the data.

We use this AM system for aggregating and analyzing the breast cancer tumor registry and EMR data sets. Cancer patient factors, namely, comorbidities and hormone-receptor status, together with diagnosis and treatment information like cancer stage, chemotherapy protocol, and secondary therapeutic drugs constitute the set of attributes. Once the AM model is generated, automated queries are run using a C# API provided by the implementation system of our choice to retrieve the associations between the drugs of interest (both chemotherapy and secondary treatment drugs) and patient factors.

**Outcome measurement:** In this paper, we investigate a specific clinical question of identifying the correlations between treatment drugs and the stage of breast cancer and hormone-receptor status of patients. Our output metric is always the number of supporting evidences, i.e., the count of patients with identical factors who are administered a particular drug. Simple graphical displays such as scatter plots are generated using Python version 2.7 to visualize the presence and strength of the correlations. Comparisons are made with SQL to highlight the usefulness of the AM system in identifying the correlations with less implementation effort and in identical level of accuracy.

Two-sample t-tests with unequal sample variances (Welch’s test) are run using R version 3.1.3 to test the statistical significance of the correlations, where the first sample consists of the proportion of patients of a particular type (e.g., specific hormone-receptor status) who are treated with a fixed set of chemotherapy or both chemotherapy and secondary drugs, and the second sample comprises the proportion of patients who are receiving the same set of drugs but are not of the first type. Patients of unknown types are excluded while computing the proportions. The null hypothesis that the two samples means are equal is rejected if the corresponding p value is less than 0.05, thereby establishing a correlation between the patient type and administered drugs with 95% significance level.

**Results**

We obtained the results on an Intel Core i7-4500U processor with 8 GB RAM and 1.8 GHz processor speed in 64-bit Windows 8 operating environment. Using the AM system, it took 6.065 minutes to generate all the patient counts as functions of breast cancer stage and hormone-receptor status, and 0.781 minutes to obtain all the patient counts as functions of hormone-receptor status and specific chemotherapy and hormone-targeted drugs. While SQL queries took about the same time to execute, they required a lot more programming effort, as shown by the snippet of required SQL query in Figure 3. No differences in the retrieved patient counts were observed between the two approaches.
public CancerDB(String host, String user, String password) {
    frm = new IAMFramework();
    // login
    sessionId = (new Random()).nextInt(1, 1 << 16);
    ActionFlags f = new ActionFlags();
    f.SessionID = sessionId;
    frm.LOGIN(host, user, password, "ManagerT", f);
    // get model
    f.GET.OpQualifiers.OpApplyVia =
        ApplyVia.enApplyViaDictionary;
    GenericList modelList = frm.GET("CancerData201504",
        null, null, new ReturnAs(), f);
    model = frm.ExtractPersistable(modelList)[0];
}

public Core.KeyValuePair getConceptByConceptName(String conceptName) {
    ActionFlags f = new ActionFlags();
    f.SessionID = sessionId;
    f.GET.OpQualifiers.OpApplyVia =
        ApplyVia.enApplyViaDictionary;
    return frm.GET(model, conceptName, null, ",", f)[0][0];
}

Figure 3: Screenshots of code using (a) associative memory system to query any combination of patient, tumor, and drug factors, and (b) SQL to specifically query patient counts for drugs based on hormone-receptor status

Figure 4: Variations in the number of breast cancer patients who various stages of invasive breast cancer and hormone-receptor status displayed using a scatterplot with circle size proportional to the patient count.

Using our AM model, we first examined the relationship between breast cancer stage and patient hormone receptor status. Figure 4 shows the results of variations in the number of patients who have different stages of invasive breast
cancer and their hormone-receptor status using a scatterplot with circle size proportional to the patient count. The results indicate a predominance of patients with ER+/PR+ and triple positive (ER+/PR+/HER2neu+) status along with early stage patients (1A, 2B, and 2C) within the extracted cohort. Similar predominance has been observed in other breast cancer patient cohorts.

Figure 5: Variations in the number of breast cancer patients receiving different chemotherapy drugs based on hormone-receptor status displayed using a scatterplot with circle size proportional to the patient count.

We next examined the correlations derived from the AM model between hormone-receptor status and choice of chemotherapy agent that was administered, which is shown in Figure 5 as a scatter plot. The figure indicates a predominance of the use of cyclophosphamide, doxorubicin, and paclitaxel, which are agents associated with recently recommended chemotherapy protocols for patients who have stages I-III breast cancer.

We then examined the relationship between hormone-receptor status and hormone-targeted treatments in the form of trastuzumab, pertuzumab, and methotrexate. Clinically, according to national guidelines such as those published by the National Comprehensive Cancer Network, we expect that the hormone-targeted treatments would be only used in patients who are HER2/neu positive. In our analyses using the AM model, we, however, found that a small subset of patients who were HER2/neu negative (categorized as triple negative or ER, PR positive) had also received hormone-targeted treatments, as shown in Figure 6.

We also computed two-sample t tests to validate the correlations between chemotherapy and hormone-targeted agents and hormone-receptor status in breast cancer patients. The results are shown in Table 1. Statistically significant differences between patients who were ER+, PR+, and ER-PR+ versus patients who had other hormone statuses in the receipt of both chemotherapy and a hormone-targeted agent were observed, which is clinically expected. Interestingly though, patients who were HER2/neu positive and ER, PR negative were not different in their receipt of chemotherapy and hormone treatment. Furthermore, HER2/neu positive patients with some combination of ER or PR positive status actually received significantly less chemotherapy and hormone-targeted agents than the others. Both of these findings contradict clinically expected patterns.
Table 1: \( p \) values corresponding to two-sample \( t \) tests to validate the correlations between drug treatment and hormone-receptor status in breast cancer patients with statistically significant correlations (\( p < 0.05 \)) highlighted in bold; all the tests are one-sided with * and ** denoting greater and lower means of the first sample, respectively.

<table>
<thead>
<tr>
<th>Hormone-receptor status</th>
<th>Drugs</th>
<th>Both chemotherapy and hormone treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chemotherapy</td>
<td></td>
</tr>
<tr>
<td>Triple negative vs. other*</td>
<td>0.19</td>
<td>0.25</td>
</tr>
<tr>
<td>ER+, PR+, and ER-PR+ vs. other*</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>HER2/neu+ vs. other*</td>
<td>0.29</td>
<td>0.10</td>
</tr>
<tr>
<td>ER-HER2/neu+, PR-HER2/neu+, triple positive vs. other**</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Discussion

The integration of heterogeneous data from multiple source systems is a time-consuming process and a rate-limiting step to rapid exploration of data, whether such efforts occur through standard data warehousing approaches or by data integration platforms, such as i2b2. In our work, we use a novel method for ingesting data from multiple sources into an associative memory (AM) model that is designed to mimic the ways humans store information through content-addressable memory and make rapid associations between the concepts. We have taken a commercially available database method for associative memory and implemented a programmatic interface that allows us to query the naturally occurring clusters generated by the model. In this paper, we have shown the feasibility of the AM approach to organize heterogeneous data efficiently based on attributes, and rapidly generate associations among the data attributes. Although the run time of corresponding SQL queries is similar to the AM model, the SQL query requires more technical expertise to formulate and thus more data engineering effort.
Using the approach, we interrogated the AM system to find associations between breast cancer stage, hormone-receptor status, and drug administration. The results of these analyses show expected clinical relationships between tumor factors and therapy choice. However, we also found a few anomalous patterns between hormone-receptor status and the use of hormone-targeted therapy. These patterns may be due to several factors including incorrect pathological identification or administration records, and thus require us to validate the information using other sources of data. Of note, however, is that our statistical validation step did not provide us the expected pattern of distinguishing hormone-receptor status between those who received chemotherapy alone and those who received those agents in conjunction with hormone-targeted therapy.

The AM approach we have chosen allows us to make rapid exploration of other patterns within the data and to easily scale the method to much larger data sets. For our future work, we will import data into the AM model from other types of cancers that are within the tumor registry, and we will begin exploring expected and unexpected clinical patterns with clinical feedback. We recognize that reviewing such highly multi-dimensional relationships in the data will require visualization approaches, and we are developing 3D methods using the OpenGL standard to enable such exploration directly by clinicians.

Acknowledgments

We thank the assistance of Drs. Tracy Onega and Judy Rees, Directors of the Norris Cotton Cancer Center Registry Resource for their support of this research.

References

1. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Medical Care, 2002; 40(8), IV-3.
The Nurse Watch: Design and Evaluation of a Smart Watch Application with Vital Sign Monitoring and Checklist Reminders

Magnus Bang, PhD¹,², Katarina Solnevik, RN³, Henrik Eriksson, PhD¹
¹Department of Computer and Information Science, Linköping University, Sweden; ²SICS East Swedish ICT AB, Sweden; ³The Intensive Care Unit, Linköping University Hospital, Region Östergötland, Sweden

Abstract

Computerized wearable devices such as smart watches will become valuable nursing tools. This paper describes a smart-watch system developed in close collaboration with a team of nurses working in a Swedish ICU. The smart-watch system provides real-time vital-sign monitoring, threshold alarms, and to-do reminders. Additionally, a Kanban board, visualized on a multitouch screen provides an overview of completed and upcoming tasks. We describe an approach to implement automated checklist systems with smart watches and discuss aspects of importance when implementing such memory and attention support. The paper is finalized with an in-development formative evaluation of the system.

Introduction

The nurse watch has been a valued nursing tool since the early 20th century. Being predominantly employed in the United Kingdom, besides offering a time reference, it provides the nurse with a quick method of taking the 60-second pulse rate of patients using a pulse scale. Due to hygiene requirements, the watch is carried up side down on a chain (fob) and attached to the uniform for fast reference. The introduction of wearable technology such as wireless computerized smart watches from manufacturers such as Apple, Motorola and Samsung provide designers with a new set of possibilities to support medical professionals. For example, these developments make it possible to design and develop specialized nurse applications that run on smart watches. However, very few smart-watch designs for nursing have been presented.

Using checklists and reminders in clinical care have been shown to provide significant improvements in patient outcome and patient safety¹. Checklists can support standardized nursing care, aid workflow, and act as reminders of work-to-do²,³,⁴. Moreover, studies show that appropriately-designed workplace arrangements and applications can offload cognitive demanding tasks and allow the clinician to focus on the important activities. Some studies, however, report that reminders have less significance to clinical outcome and routines⁵. Nevertheless, studies have identified compliance problems with both paper-based and computerized checklists⁶ such as neglecting to use a checklist or not acting on a computerized reminder.

We have developed a smart-watch application and an adjunct Kanban task board to be used by ICU nurses during day and night shifts. The basis for the design was to approach a set of issues related to the use of checklists and reminders at a Swedish ICU. The goal was to improve workflow and patient awareness at the ICU. To direct the design, we applied the theoretical framework of Distributed Cognition⁸,⁹ to understand underlying aspects of checklists and reminders that are of importance when designing memory and attention support. This paper is organized as follows: First, we review research on checklists and their usefulness. Second, we discuss theoretical aspects of checklist and reminders based on the theory of Distributed Cognition. Third, the method section describes MODUS, the framework in which we have developed our approach as well as discussing how we elicited the requirements for the system. The remainder of the paper presents the smart watch app, its formative evaluation and discusses the pros and cons of the overall approach.

Background

Nurses have many responsibilities during day and night shifts. Numerous studies have demonstrated how reminders and alerts can be helpful in focusing the nurses’ attention on specific tasks. Reminders and alarms can be seen as decision-support tools¹⁰ designed to improve patient outcomes, reduce medical errors, and increase compliance with standards of care. A variety of formats have been devised such as paper-based and computerized reminders. Rind and colleagues showed that computer-based alerts and related changes in clinical practice had a respectable effect on preventing renal impairment and function of patients¹¹. Oniki and colleagues studied computer-generated reminders on charting omissions and they could show a clear decrease in charting deficiencies¹². The mid-day reminders
appeared to reduce deficiencies in the nurses’ charting and the reminders were helpful in focusing the nurses’ attention. Thongprayoon and colleagues showed that an electronic checklist reduced significantly provider workload and task errors\(^1\). This study also suggested that electronic checklists are feasible in the ICU setting. Before we describe our approach to the design of reminders using smart watches and Kanban boards, let us discuss some theoretical aspects of checklists and reminders.

**Distributed cognition of checklists and reminders**

According to Norman, are there numerous cognitive aspects to the checklist\(^4\). First, without the list one needs to remember all tasks. On the other hand, when having a list, we need to remember little since the planning and remembering process was done in advance. At the time we enact the actions of the list, we do not need to repeat the planning and remembering process and this offloads cognition during execution. However, the use of a list also introduces new tasks: remembering to consult the list and reading and interpreting the tasks on the list. Naturally, having the physical reminder *visible and upfront* is important when relying on them for remembering tasks. According to Hutchins is this situation an example of *Distributed Cognition*; the computational task of remembering is distributed across time and across internal and external representations\(^8,9\) which comprise a robust memory system.

**Method**

**Requirements elicitation**

The requirements elicitation and formative evaluation for the Nurse Watch were done iteratively and dynamically with a team of nurses using a Scrum methodology\(^15\). This comprised software development, user demo, user feedback and subsequent redesign for four one-week sprints. Moreover, the head nurse and team redesigned and integrated a set of paper-based standard checklists and activities that were transformed to an internal XML format in the MODUS system. This transformation also comprised time stamping a set of nursing activities that previously not had been time aligned in the paper version of the checklist.

**System development**

The Nurse Watch was developed and integrated in MODUS, which is a comprehensive visualization environment and database system designed to support medical rounds and remote monitoring of patients in critical care. MODUS collects data from a set of Philips MP70 patient monitors, Braun Space Station Pumps and Maquet Servo-I ventilator systems as well as local picture and laboratory systems. This data is visualized on a motorized 46-inch multitouch tabletop during the round. Additionally, a smartphone application (Android) and Google Glass augmented-reality support remote monitoring of patients’ vital signs. The idea in this project was to add an unobtrusive new user interface to MODUS with a focus on nurses and their needs. In this project, we were able to reuse programming code from our Google Glass implementation and provide subset functionality to the watch. Moreover, we added a Kanban\(^16\) board to our multitouch tabletop to visualize all done and remaining checklist events on a timeline.

**Formative evaluation**

A formative evaluation\(^17\) was performed with seven experienced nurses and nurses’ aides of different ages and IT experience measuring *task completion* and *subjective usability* of the user interface of the Nurse Watch. All subjects had never used a smart watch before. Scenarios and eight test cases were developed covering all possible tasks and user interface components of the clock. The user evaluation was designed as a Wizard of Oz study\(^18\) that entailed simulated alarms and reminders triggered behind the scenes by our test staff. After a very brief introduction to the Nurse Watch the evaluated started and included a think-aloud methodology when the test subjects completed tasks in the scenario. After the initial test session subjects completed the SUS subjective usability questionnaire\(^19\). The session was finalized with a brief interview on the feasibility of the overall approach.

**Figure 1.** The three-tier model of MODUS allows medical equipment such as patient monitors to communicate with the Nurse Watch via the MODUS app.
Result

The current version of the Nurse Watch was developed for the Motorola 360 smart watch\textsuperscript{25} in Java for Android and connected to a set of related systems in a tree-tier model. Figure 1 shows the software architecture. In the present implementation, the Nurse Watch communicates wirelessly via Bluetooth Low Energy to an assistant smart phone also worn by the nurse. A smart phone application manages communication via Wi-Fi with the underlying patient server and databases in MODUS where vital signs data and checklist are stored. Moreover, the smart phone application allows the user to set personal alarm threshold values for the own Nurse Watch. Figure 2 shows the Motorola 360 smart watch running the Nurse Watch application.

\textit{Nurse watch design features}

The Nurse Watch provides two primary features: (1) real-time vital-sign monitoring and (2) reminders of what to do. Navigating the Nurse Watch app is done by ordinary finger sweeps and clicks on the glass surface on the watch. Since this is a very small screen, we were forced to minimize the amount of information shown. Moreover, for patient-safety reasons, we decided to have the patient name and id highly visible on most screens even though this takes up screen surface. The reason for this decision was to decrease the risk of introducing mode errors in terms of assessing the wrong patient based on information from the watch. Figure 3 shows the different user interface panes for vital-sign monitoring and reminders. Let us first discuss the vital sign monitoring part of the Nurse Watch.

The patient pane with vital signs is the primary home screen (see Figure 3, left). This screen is dynamically updated every other second. Scrolling up and down with the finger changes which patient is shown. What is presented is a subset of available patient parameters in the MODUS environment. To create a set of vital signs to be shown on the watch, the clinician chooses these parameters on a digital desk or PC using a drag and drop methodology. This activity allows the parameters to be sent and shown on the smartphone and on the watch. Additionally, the nurse can, using the adjunct phone app, set suitable alarm thresholds for the watch. Figure 4 shows the smartphone app for setting the alarm thresholds.

We implemented a two-level notification scheme in the Nurse Watch: vibration alarms and visual cues. Moreover, we decided to have two types of vibration alarms to notify the clinician on the level of urgency. Two consecutive vibrations denote an urgent alarm and indicate a violation of the threshold of a vital sign. A longer vibration is related to upcoming items on the checklist. Moreover, to clearly indicate what type of vital sign that was out of bounds, the color of the vital sign text changes to red and is flashing to make it stand out.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{The Nurse Watch showing the breath rate of a patient. The additional vital parameters of the patient are automatically shown.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{The user interface components of the Nurse Watch application. The app provides real-time monitoring of patients’ vital signs and deliver reminders on upcoming task. The nurse can classify a task as done (Swe., Avklarad) or postpone it (Swe., Skjut upp).}
\end{figure}
The Nurse Watch provides notifications on upcoming tasks. These are generated automatically in MODUS based on a time-stamped checklist in XML format that is executed in real time. As mentioned, the system generates a second level notification in terms of a longer vibration in the watch. The nurse can now directly view the task and classify it as completed or postpone it. The status of the task is then updated on the watch and on the Kanban board in terms of visual cues.

The Kanban board

Originally, Kanban boards were introduced in the manufacturing industry to support workflow and to provide an overview of the current work situation. In this project, we combined the traditional checklist with Kanban visualization. The idea is straightforward: every task on a checklist is time-stamped and visualized on a Kanban board. We developed a special pane for this in the MODUS system. The process is initialized when the patient is admitted to the ICU. Then, the head nurse chooses among a standard set of care plans/checklists in MODUS and these are shown on the Kanban board and are sent to the watch. Figure 5 shows a checklist visualized on the Kanban board. In MODUS these tasks time-activated, which triggers the Nurse Watch of the responsible clinician.

Formative evaluation

The formative in-development evaluation revealed some user interface omissions in the first version of the nurse app prototype. The task completion rate for the eight primary tasks were 83 percent and this clearly indicated that there were problems with parts of the user interface. Primarily, two user interface problems were identified; glitches that were resolved in the finalized version of the Nurse Watch. One problem was an inadequate implementation of the alarm screen that required subjects to remember alarm types and patient names during navigation. A better design would have notified the nurse and allowed direct access to the corresponding patient and parameter screen. The second user interface issue was related to insiciencies in how we displayed tasks as being done or not.

The SUS subjective usability scale measures several aspects of the usability of IT systems such as the need for support, training, and complexity of the user interface. The SUS mean score for the first Nurse Watch prototype was 70, which is regarded as OK but not exceptional. Particularly low scored a question on the comfort and security of the prototype. We believe that this result was due to the above discussed user interface issues combined with a very brief introduction to the new technology.

During the interviews the clinicians were generally positive towards the Nurse Watch and the overall approach. The nurses expressed that the user interface of the watch was easy to use and understand. The vital sign monitoring and the related threshold alarms were, by the lion part of the clinicians, seen as the most useful feature of the watch. Checklist reminders were also regarded as a useful function. However, a concern was on the relevancy of the reminder. One nurse stated, for example, that for it to be really useful, it should prompt only key medical tasks, not ordinary standard care task. Moreover, a suggestion was to include functionality to create own reminders that are displayed on the watch during the shift. All nurses expressed concerns regarding hygiene requirements and discussed how it could be worn to comply with these necessities. This discussion also included worries that the watch might reveal information to outsiders. Solutions included screen savers that blacken the display and ways to wear it on a chain (fob) in the pocket. One nurse expressed the need to have a tablet computer with the above-discussed features rather than a smart watch.

Figure 4. Part of the MODUS smart phone app that let the clinician set alarm thresholds for the Nurse Watch.
Discussion

Smart watches can become valuable tools in the healthcare environment. We have exemplified two primary features that can be useful – remote patient monitoring and reminders – that could improve situation awareness and support the memory of medical professionals. On the positive side is the unobtrusiveness of the device that might be its principal advantage in the healthcare environment, where nurses don’t necessarily want to carry bulky tablets or smartphones. On the negative side is screen size, which limits it usefulness as an interactive device. Moreover, energy consumption is still a problem even thought Bluetooth Low Energy is used for communication. Hence, a priority in development should be on applications that make the best out of the energy available on the platform. Related to the above is the reliability of the radio communication. This is still an issue and must be investigated further.

Many healthcare organizations have banned wristwatches because they have been identified being a transfer vector of healthcare associated infections. A possible approach to minimize this risk is to wear it with a chain hanging from the blouse or coat. However, this might impair user interaction. Software developers must account for how clinicians want to wear the device in order to find a suitable interaction scheme. Nevertheless, it needs to be washable, resist disinfecants, and have a liquid ingress protection reaching IEC 60529 level IP67.

Having automated data capture at the bedside with real-time patient monitoring feeds to devices such as smart watches introduce a set of general problems related to control and situation awareness. Traditionally, ICU nurses have the role of filling out the patient chart to record the vitals of the patient manually. Nowadays automated patient monitoring and recording systems, which risk setting the clinician out of the information and control loop, is increasingly taking this role. Therefore, vendors are requiring that the nurses in the ICU environment validate automatically collected data coming from the patient monitors. This problem also applies to the to smart watches since an increased reliance on automated feeds may inflict on the formation on situation awareness. Hence, there is a difficult tradeoff that needs to be assessed between information feeds and direct information from the patient at the bedside.

Many studies have exemplified how clinicians create their own cognitive scaffolds as part of their practice\textsuperscript{21,22}. Common examples are the ubiquitous sticky notes that can be seen as external memories of work-to-do. Smart watch implementations should support flexibility and such cognitive practices. A problem here is that the smart watch is primarily an output device, which hamper the creation of small notes. A likely design solution is to create these notes with smart phones or tablets computers and use the watch for the reminder feeds.

A risk when introducing technologies that provide alarms is alarm fatigue\textsuperscript{23}. This is particularly apparent in the case of wearable technologies such as smart watches. For the Nurse Watch, we developed different alarms levels. This is probably only part of a solution to this problem. Since these are new devices, more research needs to be conducted on how smart watches affect alarm fatigue.

Conclusion

This paper presented a design and an evaluation of a smart watch platform for ICU nurses. Appropriately designed applications running on these unobtrusive devices can become important nursing information tools in the future. We exemplified two primary features – vital sign monitoring and work-to-do reminders – that were perceived by nurses to be useful in an ICU environment. Future work includes improvements to our alarm scheme and means for nurses.

Figure 5. The Kanban board and timeline of MODUS provides an overview of the checklist. The shown checklist items are intended to remind nurses to wash the patient regularly. Each upcoming task is also sent to the Nurse Watch as a notification. On the right side is two completed tasks visualized.
to create their own reminders to support their personal memory practices. Full in situ studies are required to appropriately assess the feasibility and usefulness of smart watches in the clinical environment.

References
Utilizing Multidimensional Computer Adaptive Testing to Mitigate Burden With Patient Reported Outcomes
Michael Bass, MS1, Scott Morris, PhD2, Richard Neapolitan, Ph.D1
1Northwestern University, Chicago, Il, 2Illinois Institute of Technology, Chicago, Il

Abstract

Utilization of patient-reported outcomes measures (PROs) had been limited by the lack of psychometrically sound measures scored in real-time. The Patient Reported Outcomes Measurement Information System (PROMIS) initiative developed a broad array of high-quality PRO measures. Towards reducing the number of items administered in measuring PROs, PROMIS employs Item Response Theory (IRT) and Computer Adaptive Testing (CAT). By only administering questions targeted to the subject’s trait level, CAT has cut testing times in half1. The IRT/CAT implementation in PROMIS is unidimensional in that there is a separate set of questions administered for each measured trait. However, there are often correlations among traits. Multidimensional IRT (MIRT) and multidimensional CAT (MCAT) provide items concerning several correlated traits, and should ameliorate patient burden. We developed an MIRT model using existing PROMIS item banks for depression and anxiety, developed MCAT software, and compared the efficiency of the MCAT approach to the unidimensional approach.

Note: Research reported in this publication was supported in part by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011962.

Introduction

There is increasing pressure for integration of the patient’s perspective in clinical research. An indication of this growth is the recent establishment of the Patient-Centered Outcomes Research Institute (PCORI) by Congress. PCORI’s goals include identifying the best available evidence to help patients and their healthcare providers make more informed decisions. Patient-reported outcomes (PROs) offer the patient’s perspective. The notion of the pre-eminence of the patient perspective has been prominent in literature for at least 20 years5-7. The research literature contains ample evidence for not relying solely on clinician reporting of patients’ subjective experiences. Discordance between patient and clinician reports of Health-Related Quality of Life (HRQL) and health status has been widely reported8,9. In general, clinicians have reported fewer symptoms10-13 and lower symptom severity14,15 compared to patient reports.

There is also increasing evidence that PROs can provide more information about toxicity and symptoms than physician or adverse event (AE) reports16,17. Studies have shown that, for some subjective symptoms (e.g., in sexual function), clinician ratings are no better than chance in predicting PROs of the same symptoms18,19. These findings support the need for tools to collect patient-reported symptoms and HRQL. The inclusion of PRO measures of AEs and other symptoms in clinical trials and other research recently has been recommended as a means to improve the quality and completeness of data, provide a more comprehensive reflection of patient experiences, and improve the operational efficiency of conducting studies20-22.

The need to reduce patient burden in measuring PROs has long been a concern in health outcomes research. This concern has led researchers away from reliance on classical testing models and toward item response theory (IRT) and computer adaptive testing (CAT). These methods have been used successfully for decades in education, licensure, personality assessment, and selection of military personnel. By only administering questions targeted to the individual’s ability or trait level, testing times have on average been cut in half while at the same time improving overall test reliability1. The efficiencies long noted in educational testing and credentialing are equally applicable in the measurement of PROs. Underlying every CAT is a bank of items “calibrated” to an IRT model. This model allows scores estimated from any subset of the entire bank to be expressed in the same metric. Therefore, although different individuals take different subsets of items, their scores are directly comparable23. In a CAT-based measure, an initial item from the bank is administered. A person’s response to that item provides a rough estimate of the person’s level of the trait being measured. For example, a patient who initially chooses the response, “Very Much,” when asked, “How much did pain interfere with your day to day activities?” is likely has a high level of pain interference24. The items subsequently administered would be those known to be effective in discriminating among people experiencing high pain. On the other hand, an individual who responded “Never” to this question would be asked additional questions known to be good at discriminating among people with low levels of pain interference. Because the CAT only administers the items that are most informative for each individual, patients can typically
complete a PRO CAT in an average of five questions with reliability similar to a traditional 15–25-item survey measure\textsuperscript{23}. Reducing the number of items on each instrument lowers the burden on patients and allows the measurement of more PROs, thereby supporting more comprehensive assessments.

The Patient Reported Outcomes Measurement Information System (PROMIS) is an initiative, funded by the National Institutes of Health (NIH) Common Fund, to improve assessment of self-reported symptoms and other health-related quality of life domains. CAT-based PROs have become much more accessible through PROMIS. A primary goal for PROMIS was the creation of CATs to enable efficient, psychometrically robust assessment of PROs for a wide range of health outcomes research\textsuperscript{25}. To achieve this goal, customized banks of items for measuring PROs were developed and tested at 12 primary research and data collection sites. Initially PROMIS developed and validated 27 adult and 9 pediatric item banks across the domains of physical, mental and social health. Many PROMIS instruments were developed by scientists who had previously authored well-known PROs including the Short Form-36, the Functional Assessment of Chronic Illness Therapy, and the Health Assessment Questionnaire. In practice, PROMIS banks have shown greater reliability, fewer ceiling and floor effects and more responsiveness to change than the “gold standard” instruments\textsuperscript{25}.

The PROMIS initiative lays the foundation for widespread health outcomes screening to take place in the waiting room of healthcare providers. These screenings can coincide with the current activities that usually occur while a patient is waiting for their healthcare provider, such as filling out or updating their past medical history and insurance forms. The data collected from these screenings can be incorporated into the patient’s electronic health record (EHR) alongside other clinical patient data. This will give healthcare providers a more accurate picture of their patients overall health and healthcare needs. Healthcare providers can make more informed decisions with regards to mental health referrals or treatment options.

In order to realize these benefits, PRO screenings must capture maximum information with minimum patient burden. PROMIS is unidimensional in that there is a separate item bank and therefore a separate set of questions administered for each measured trait. This has been a good initial step in addressing the issue of patient burden, but in the real world things are not always as simple as the one dimensional approach; that is, there are correlations between many of the traits assessed by the PROMIS system. For example, depression and anxiety are closely related\textsuperscript{37}, which means many questions that ascertain level of depression can also provide information about the level of anxiety. A multidimensional IRT (MIRT) model represents items concerning several correlated traits. A multidimensional CAT (MCAT)\textsuperscript{26-29} selects items from an item pool such that the items selected maximize information provided on several correlated traits. Such a model would significantly reduce the number of items administered in measuring PROs and thereby ameliorate the patient’s burden of answering many questions.

Researchers have recently realized the acute value of the MCAT approach in health services\textsuperscript{30-32}. In this research we investigated whether MCAT can lessen patient burden relative to unidimensional CAT (UCAT) in the domain of PROs. We performed this investigation using the related traits “anxiety” and “depression.” First, we developed an MIRT model for these two traits using existing PROMIS item banks. Then we developed MCAT software to administer questions based on this MIRT model. We compared the performance of this MCAT approach to the unidimensional approach currently used by PROMIS using 20,000 simulated individuals. We found that the MCAT consistently provided comparable estimates of anxiety and depression with fewer items, relative to administering separate UCATs.

**IRT and CAT Review**

Since the methodology concerns IRT and CAT, we first briefly review these two.

**Item Response Theory.** In item response theory (IRT)\textsuperscript{31,34}, we estimate an individual’s traits from the individual’s responses. Each trait is a skill, knowledge, or quality in some domain. For example, a trait may be “arithmetic ability,” while the responses are answers to arithmetic problems. Traits are hidden variables; responses are observed variables. Figure 1 shows a Bayesian network (BN)\textsuperscript{35} representing an IRT model with one trait $\theta$ (unidimensional) and five items.

The probability distribution of correctly answering a dichotomous item given the level of a trait is commonly modeled using the following two parameter logistic evidence model:

$$P(I_j = \text{Right}) = \frac{1}{1 + e^{-a_j(\theta-b_j)}},$$  \hspace{1cm} (1)

321
where $b_j$ measures the location of the item on the latent trait continuum (e.g., whether it is easy or difficult), and $a_j$ is a discrimination parameter reflecting how sensitive the item is to differences in the latent trait.

In the case of PROs there is no “correct” response. Instead, a polytomous IRT model is used to represent the probability of each option on the response scale. The probability distribution of a polytomous item with $m + 1$ categories given trait $\theta$ in the graded response model \(^3^6\) is computed as follows:

$$
\begin{align*}
P_{j0}(\theta) &= P(I_j \geq 0) = 1 \\
P_{jk}(\theta) &= \frac{1}{1 + e^{-a_j(b_j - b_k)}} = P(I_j \geq k), \quad k = 1, 2, \ldots, m \\
P_{j,m+1}(\theta) &= P(I_j \geq m + 1) = 0
\end{align*}
$$

where $b_{j1} < b_{j2} < \cdots < b_{jm}$. The probability of selecting option $k$, $k = 0, 2, \ldots, m$, is given by

$$
P(I_j = k) = P(I_j \geq k) - P(I_j \geq k + 1).$$

Traits in a given domain can be related, which means many items that ascertain the level of one trait can also provide information about the level of other traits. For example, many items that ascertain depression level can also provide information about anxiety level \(^3^7\). These relationships can be modeled using an MIRT model. In the case of MIRT, Equation 1 for the two parameter dichotomous logistic model is as follows:

$$
P(I_j = \text{Right}) = \frac{1}{1 + e^{-a_j(\theta - b_j)}}
$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ is a vector of traits and $a_j$ is a vector of discrimination parameters.

**Computer Adaptive Testing.** Computer adaptive testing (CAT) \(^3^8\) improves test measurement quality and efficiency by striving to administer optimal items to each examinee. CAT selects items sequentially to minimize the standard error of the estimate of $\theta$. Commonly employed item selection criteria used in CAT include maximizing Fisher’s information (FI) \(^3^9\), and using the Kullback-Leibler information \(^4^0\). Chen et al. \(^4^1\)
In unidimensional IRT, Fisher’s information for a polytomous item given trait level $\theta$ is as follows:

$$I_j(\theta) = a_j^2 \sum_{k=0}^{m} \left[ \frac{P_j(\theta)(1 - P_{j,k}(\theta)) - P_{j,k+1}(\theta)(1 - P_{j,k+1}(\theta))}{P_{j,k}(\theta) - P_{j,k+1}(\theta)} \right].$$

As indicated in this expression, information is a function of the individual’s standing on the latent variable $\theta$, which is unknown. To address the uncertainty inherent in the estimate of $\theta$, we computed the maximum posterior weighted information (MPWI), where the expected value of $I_j(\hat{\theta})$ is taken across the posterior distribution of $\theta$, which is unknown. To address the uncertainty inherent in the estimate of $\theta$, we computed the maximum posterior weighted information (MPWI), where the expected value of $I_j(\hat{\theta})$ is taken across the posterior distribution of $\theta$. The next item chosen is the one that maximizes the MPWI.

In the case of MIRT, FI takes the form of a matrix, and the determinant of the information matrix serves as a summary measure of the information regarding the set of traits. Thus, the next item is chosen that maximizes the posterior weighted determinant of the information.

**Methods**

The current study presents the result of a Monte Carlo simulation designed to evaluate the reduction in test length when using multidimensional versus unidimensional CAT. In order to represent a typical PRO assessment situation, we simulated responses to the actual PROMIS Anxiety and Depression banks. The PROMIS banks were developed using unidimensional IRT analyses. Therefore, an initial step toward building the simulation was to estimate the item parameters of a 2-dimensional MIRT model.

**Measures.** The PROMIS item banks were developed though a mixture of qualitative and quantitative methods. Rigorous scale construction practices were followed, with substantial input from domain experts, as well as revision based on respondent feedback. Statistical item analyses were conducted using both classical and IRT-based methods to identify the best items for inclusion in the final bank. Cella et al. The banks consist of 29 anxiety items and 28 depression items from the current PROMIS item banks.

For these items, respondents are asked to report the frequency of experiences during the past seven days. Examples of item content are as follows: “I was easily startled” (anxiety), and “I felt worthless” (depression). Responses were indicated on a 5-point scale (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always).

**Estimating the MIRT Model.** Calibrating an IRT model involves obtaining estimates of item parameters described in Equation 2. Unidimensional item parameters were available from prior work on the development of the PROMIS banks. For the MCAT, a preliminary analysis was conducted to obtain the necessary item parameters. Item parameters for a 2-dimensional Graded Response Model were estimated using the program IRTPRO.

This step utilized existing data collected for the development of PROMIS instruments. The sample was 52% female with a median age of 50. The racial composition was 82% white, 9% black, 8% multi-racial, and 1% other. Nine percent were Latino/Hispanic. The current study used a subsample of 7945 adults sampled from the general population who completed the anxiety and depression items.

The parameter estimates from the MIRT model were found to correspond closely to those from the unidimensional analyses (all correlations > .9). Items tended to have large discrimination parameters, indicating that each item was a strong indicator of the underlying trait. The average a parameter value was 2.78 (ranging from 1.41 to 3.76) for Anxiety and 3.23 (ranging from 2.21 to 4.72 for Depression).

Item location (b) parameters indicated that the items tended to reflect higher levels of anxiety and depression. For example, among individuals with a slightly below average level of the trait (0.5 SD below the mean), the most common response selected was the lowest option on the 5-point scale. This indicates that the items in this bank are primarily sensitive to moderate to high levels of anxiety and depression, and are less able to distinguish among individuals with low levels of anxiety and depression.

The preliminary analysis also yielded estimates of Anxiety and Depression levels for each individual in the data. The distribution of both anxiety and depression were positive skewed, with approximately 70% of individuals...
obtaining trait values near the mean ($\theta = 0$) or slightly below ($\theta = -1$), and the remaining 30% demonstrating higher levels of the trait. The two traits were highly correlated, $r = 0.88$, $p < 0.001$.

**Monte Carlo Simulation.** The simulation consisted of generating data for 20,000 respondents who completed both the UCAT and MCAT versions of the PROMIS Anxiety and Depression measures.

We first generated the true trait levels of each respondent by sampling from the bivariate frequency distribution obtained during the IRT calibration step. This approach produced simulated data representative of anxiety and depression levels observed in the general population.

Each simulated respondent was administered three CATs: MCAT for anxiety and depression, UCAT for anxiety, and UCAT for depression. First, an initial item was selected by the CAT algorithm. Next, a probability distribution for the individual was generated according to that individual’s true trait levels and the IRT model for that item. The IRT model corresponding to each CAT (i.e., unidimensional or multidimensional) was used to generate a probability distribution. Next, a response was generated according to this probability distribution.

After each item, the CAT then updated the posterior distribution of $\theta$ based on the individual’s response, along with the expected a posteriori (EAP) estimate $\hat{\theta}$, and the standard error (SE) of this estimate. This updated theta distribution was used by the CAT algorithm to select the next item. This was repeated for the first 20 items administered by each of the CATs.

The CAT was designed to stop when the SE of each $\theta$ estimate falls below a threshold. The SE is a measure of uncertainty in the trait estimate (i.e., how far the estimate is expected to fall from the true trait level). Consequently, when the SE falls below the threshold, this indicates that the trait has been estimated to an acceptable level of precision, and is unlikely to change substantially on further testing. For the current study, we considered a variety of thresholds, ranging from 0.2 to 0.4 in increments of .025. This range was centered on the threshold of 0.3 currently implemented in the unidimensional PROMIS CATs.

Two primary outcomes of interest were the number of items required to achieve the stopping rule, and the accuracy of the resulting EAP estimates, operationalized as the root mean square error,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}[(\hat{\theta}_{Anx,i} - \theta_{Anx,i})^2 + (\hat{\theta}_{Dep,i} - \theta_{Dep,i})^2]}{N}},$$

where $\hat{\theta}_{Anx,i}$ and $\theta_{Anx,i}$ represent the EAP estimate and true value of trait $t$ for respondent $i$, and $N$ is the number of simulated respondents.

**Table 1.** Percent of simulated respondents meeting the stopping rule on both anxiety and depression within 20 items.

<table>
<thead>
<tr>
<th>SE Cutoff</th>
<th>MCAT</th>
<th>UCAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>56%</td>
<td>63%</td>
</tr>
<tr>
<td>0.225</td>
<td>63%</td>
<td>64%</td>
</tr>
<tr>
<td>0.25</td>
<td>67%</td>
<td>67%</td>
</tr>
<tr>
<td>0.275</td>
<td>71%</td>
<td>72%</td>
</tr>
<tr>
<td>0.3</td>
<td>79%</td>
<td>76%</td>
</tr>
<tr>
<td>0.325</td>
<td>82%</td>
<td>80%</td>
</tr>
<tr>
<td>0.35</td>
<td>85%</td>
<td>86%</td>
</tr>
<tr>
<td>0.375</td>
<td>88%</td>
<td>89%</td>
</tr>
<tr>
<td>0.4</td>
<td>91%</td>
<td>90%</td>
</tr>
</tbody>
</table>
Results

An initial analysis indicated that not all simulated respondents reached the stopping rule within the 20 simulated items (see Table 1). This was particularly problematic for the more stringent stopping rules, which required a SE below 0.3. Further analyses revealed that the cases that failed to meet the stopping rule generally occurred when one or both of the traits were less than 0. As noted above, the items available in the Anxiety and Depression banks tend to be more sensitive to higher levels of these traits. The absence of items targeted toward lower levels of these traits limits the precision obtainable for individuals with below-average anxiety or depression. Thus, the administration of additional items is unlikely to change this result.

Because the outcome measures could be computed only for cases that met the stopping rule, respondents who did not satisfy this requirement were excluded from subsequent analyses. To ensure comparability, a respondent was included in the results only when the stopping rule was reached for both the UCAT and MCAT for a particular cutoff. However, the number of examinees was allowed to vary across cutoffs, so that more data was available for less stringent cutoffs (e.g., N=16,817 for SE<0.4) than for more stringent cutoffs (e.g., N=11,262 for SE < 0.2).

The average number of items required to reach the stopping rule is presented in Table 2. The MCAT required fewer items at each of the cutoff levels (differences were all significant, p<0.001). On average the MCAT required 1.2 fewer item than the UCAT. Furthermore, despite the reduced number of items, the MCAT and UCAT resulted in equally accurate trait estimates, as indicated by similar RMSE values. The greater efficiency of the MCAT is illustrated in Figure 2, which plots the RMSE against the average number of items required for each stopping rule. The lower curve for the MCAT represents greater accuracy given the same number of items, relative to the UCAT.

Table 2. Number of Items Required to Reach the Stopping Rule on Both Traits and RMSE of UCAT and MCAT.

<table>
<thead>
<tr>
<th>SE Cutoff</th>
<th>N</th>
<th># Items to Cutoff</th>
<th>UCAT</th>
<th>MCAT</th>
<th>UCAT</th>
<th>MCAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>11262</td>
<td>16.3</td>
<td>14.9</td>
<td>0.27</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>0.225</td>
<td>12454</td>
<td>12.6</td>
<td>11.6</td>
<td>0.30</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>12727</td>
<td>10.1</td>
<td>9.3</td>
<td>0.33</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>0.275</td>
<td>13137</td>
<td>8.8</td>
<td>7.6</td>
<td>0.35</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>13910</td>
<td>7.7</td>
<td>6.6</td>
<td>0.39</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>0.325</td>
<td>14565</td>
<td>7.4</td>
<td>5.7</td>
<td>0.39</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>15489</td>
<td>6.9</td>
<td>5.3</td>
<td>0.41</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>0.375</td>
<td>16343</td>
<td>6.2</td>
<td>5.2</td>
<td>0.45</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>16817</td>
<td>5.9</td>
<td>4.8</td>
<td>0.46</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>9.1</td>
<td>7.9</td>
<td>0.37</td>
<td>0.37</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion

The results of the Monte Carlo simulation clearly demonstrate the benefits of multidimensional adaptive testing for patient reported outcomes. Consistent with prior research\textsuperscript{28}, MCAT was found to produce accuracy similar to UCAT while reducing the number of items administered.

The absolute reduction in test length was considerable. The UCAT required 9 items on the average, while the MCAT required only 8 items. This corresponds to an 11\% reduction of the test length, which represents an important reduction in patient burden. However, the reduction in test length observed here is less that has been found in other MCAT applications\textsuperscript{28}. This may be due to the nature of the items examined in study. In general, each item provides considerable information about the trait. The small number of items needed to obtain precise trait estimates may limit the potential for refinements such as MCAT to reduce test length to the same degree as that achieved when there are a large number of items.

A strength of the current simulation is the use of actual items from the widely used PROMIS anxiety and depression banks to build the simulation, thereby enhancing the realism of the simulation. Nevertheless, the results represent only these two traits, and may not generalize to other PRO measures.

A distinctive characteristic of anxiety and depression is their very high correlation ($r=0.88$). It is possible that MCAT will demonstrate less benefit over UCAT when less correlated traits are modeled, because each trait will provide less information about the other. On the other hand, EAP trait estimates tend to be slightly biased toward the mean of the prior distribution, and this effect is heightened in MCAT when the traits are highly correlated. Therefore, MCAT with less correlated traits may exhibit less bias and greater accuracy. Further research is needed to replicate these findings with other combinations of traits.

The sample of participants for the current study was mostly white (82\%), and the generalizability of these findings across ethnic groups merits further study. Mean differences across groups should not affect the performance of MCAT assessments. However, if ethnic subgroups interpret and respond to items differently, this will affect the parameters of the IRT model, and consequently the accuracy of the IRT-based adaptive tests. Although there is evidence that measures of depression may function differentially across ethnic groups\textsuperscript{43,44}, little is known about the impact of these differences on UCAT and MCAT measures.

References

A Low-Cost Method for Multiple Disease Prediction

Mohsen Bayati, PhD\(^1\), Sonia Bhaskar, MS\(^2\), and Andrea Montanari, PhD\(^2,3\)

\(^1\)Graduate School of Business, Stanford University
\(^2\)Department of Electrical Engineering, Stanford University
\(^3\)Department of Statistics, Stanford University

Abstract

Recently, in response to the rising costs of healthcare services, employers that are financially responsible for the healthcare costs of their workforce have been investing in health improvement programs for their employees. A main objective of these so-called “wellness programs” is to reduce the incidence of chronic illnesses such as cardiovascular disease, cancer, diabetes, and obesity, with the goal of reducing future medical costs. The majority of these wellness programs include an annual screening to detect individuals with the highest risk of developing chronic disease. Once these individuals are identified, the company can invest in interventions to reduce the risk of those individuals. However, capturing many biomarkers per employee creates a costly screening procedure. We propose a statistical data-driven method to address this challenge by minimizing the number of biomarkers in the screening procedure while maximizing the predictive power over a broad spectrum of diseases. Our solution uses multi-task learning and group dimensionality reduction from machine learning and statistics. We provide empirical validation of the proposed solution using data from two different electronic medical records systems, with comparisons to a statistical benchmark.

Introduction

Early identification of individuals who are at a higher risk of developing a chronic disease is of significant clinical value, as it creates opportunities for slowing down or even reversing the pace of the disease [1]. In particular, 80% of the cases of cardiovascular disease and diabetes, and 40% of the cancer cases can be treated successfully at an early stage [2], preventing the need for expensive medical procedures due to complications occurring during the standard care of chronically ill patients. This early stage intervention is also of significant cost value, as avoiding these expensive procedures and associated costs of care leads to a reduction in healthcare spending. In 2012, nearly 76% of Medicare spending was on patients, comprising just 17% of the Medicare population, with five or more chronic diseases: Alzheimer’s disease, cardiovascular disease, diabetes, renal disease, and lung disease [3].

Recently, one solution to this rising cost burden has been the introduction of incentive programs by an increasing number of self-insured companies, i.e., companies who pay for healthcare costs of their workforce. In 2013, 86% of all employers, which include Johnson & Johnson, Caterpillar, and Safeway, offered such incentivized programs. These programs involve, among many other aspects, incentivizing employees to undergo a health risk assessment (HRA) in order to identify employees with risk factors for major chronic diseases. The HRA typically involves collecting basic lab results like lipid panels, or other relevant biomarkers such as blood pressure, age, gender, height and weight. This data about the employee is then used to calculate risk factors based on simple clinical rules (see [4] for a list such rules). However, these simple clinical rules suffer from poor prediction accuracy since they rely on few biomarkers and do not account for risk variations in different populations, resulting in reduced efficacy of the program and undermining its objectives, suggesting the need for more effective methods for multiple disease screening and early identification. In particular, a false positive (mistakenly assigning high risk to a healthy patient) leads to unnecessary intervention costs while a false negative (mistakenly assigning low risk to a high risk patient) would be a lost opportunity to avert a large future healthcare bill to the employer, and it has been shown that higher predictive accuracy directly improves the cost-effectiveness of such programs [5].
Statistical data-driven methods present one principled way to obtain a set of biomarkers for such an assessment which provides high disease prediction accuracy. The use of such methods for disease prediction has grown rapidly over the past few years. An admittedly incomplete list of studies include: stroke prediction via Cox proportional hazard models [6], relational functional gradient boosting in myocardial infarction prediction [7], support vector machines and naïve Bayes classifiers for cancer prediction [8], neural networks for mortality prediction [9], time series in infant mortality [10], cardiac syndromes [11] or infectious disease prediction [12], and dimensionality reduction for unstructured clinical text [13]. However, these past works have focused on individual disease prediction, rather than determining a set of biomarkers which are jointly predictive of several diseases. While it is possible to combine the relevant biomarkers selected by each individual disease prediction model, this leads to a larger set of biomarkers than is necessary, and we will later show this is less cost effective than our proposed method.

In this paper, we present a method which relies on multi-task learning [14] model and group regularization [15]. The multitask learning model we consider allows for determining a small set of biomarkers which are optimal for prediction of multiple diseases, and is hence low cost. The model acts by learning the relevant biomarkers for multiple diseases simultaneously, while forcing these relevant biomarkers to be sparse [14, 16, 17], resulting in a lower cost model. Furthermore, it is known to be highly generalizeable and thus to adapt well to new data; we refer the reader to [18] for a detailed survey.

While these aforementioned works share methodology with our proposed method, to the best of our knowledge none have not yet addressed a setting where one desires to build a small, and thus low cost, universal set of features (biomarkers in our case) that can be used to predict multiple diseases. The work of [19] also considers multitask learning in the context of logistic regression and multiple disease prediction, however, their model does not enforce the disease prediction tasks to share the same set of biomarkers. Their model constructs a shared set of latent predictors as linear combinations of groups of features (diagnosis history) and a $\ell_1$ regularizer is used to ensure that each individual disease prediction relies on a small set of these predictors. Hence, their approach does not enforce a shared group of biomarkers to be used as predictors for all diseases, which makes their method less suitable for our purpose. Our model also uses the raw biomarkers as the predictors, which is more interpretable and enables our model to directly ensure that the set of biomarkers is small. Thus we present our proposed method as a solution to obtaining an accurate, low cost disease screening methodology. We evaluate our method on two patient populations, and compare the cost of our method to the current statistical benchmark, individual disease prediction. We find that our method has comparable, and in some cases improved, accuracy at much lower cost.

1 Methods

1.1 Patient Data

We considered two patient populations for developing our prediction tool, the Kaggle Practice Fusion dataset [20], which is publicly available, and the patient records from Stanford Hospital and Clinics, which will be referred to as the SHC dataset throughout 1. The SHC dataset consisted of 73,842 patients, and 1,313 possible laboratory exams; while the Kaggle dataset consisted of 1,096 patients with 285 possible exams; a comparison summary is given in Table 2.

The patient data from the SHC dataset was taken as follows. For each disease $j$ ($j = 1, \ldots, K$, with $K = 9$ for the SHC dataset), we collected the biomarkers of patients during a one-month period, which occurred during the year 2010 to 2011. These biomarkers included quantities such as age, gender, and the results of laboratory exams such as hemoglobin A1C, cholesterol, as well as blood cell counts2. If there were multiple laboratory exam results during this month, they were averaged. If patients had results over multiple months, we took the first month for which results occurred. If the patient was diagnosed with disease $j$ during this month, we removed the patient from

---

1 Due to HIPAA privacy restrictions, this dataset cannot be made publicly available.

2 However, we did not have access to biomarkers such as blood pressure, height, and weight.
the dataset, i.e., we only considered patients who had never had disease $j$ before the end of the month in which their results were taken. We collected these biomarkers into a vector $X_i$, $i = 1, \ldots, n$, where $X_{ij}$ took the value of the result of biomarker $j$ for patient $i$, and where $n$ was the total number of patients collected. We assigned each patient $i$ a label $y_i^{(j)}$ for each disease $j$, which we defined to be $+1$ if the patient has a positive diagnosis of disease $j$ within one to thirteen months, beginning the month after biomarkers were collected, and $-1$ otherwise. Lastly, we took the intersection over all diseases for which we collected patients. For the SHC dataset, we noticed that the set of patient-months for all diseases had a large intersection containing 75,619 patients, as shown in Table 2. Two of these diseases had an extremely low incidence in the Kaggle dataset, and were not studied when evaluating on the Kaggle dataset. These diseases are highlighted in Table 1 with an asterisk (*). The diseases we studied were obtained from [21, 22] and are shown in Table 1, with their formal names and abbreviations which will be used throughout this paper.

In the Kaggle dataset, due to lack of access to the exact time stamp for the lab results or diagnosis codes, the feature vectors $X_i$ were formed for each patient from the laboratory results and age information, without any use of temporal information. The label $y_i^{(j)}$ was determined by checking whether the patient had a positive diagnosis for disease $j$ in the system. Due to the small size of the Kaggle dataset, the bootstrap method was used to estimate the average and standard error of the AUC values.

For both datasets, the patients did not have results for all of the possible laboratory exams. In particular, the matrix $X$ has $98\%$, and $93\%$ missing values in SHC and Kaggle data respectively. To deal with the missing data entries, we used the standard method of mean imputation. We gave each the average value of the laboratory exam, where the average was taken over all patients who had taken the exam. We also repeated our analysis on Kaggle data set using multiple imputation (MI) with 100 imputations and the results were consistent. In particular, the relative accuracy of all algorithms and the sparsity of the models did not change. Finally, we standardized the observations so that the values of each biomarker have zero mean and variance one across the observations, which is a standard preprocessing step for many statistical methods.

### 1.2 Cost Data

Data for the cost analysis was taken from [23], maintained by Health One, Incorporated. Health One, Incorporated offers laboratory exams and has physical locations in all states in United States. Health One Lab’s website provides

---

3 We only restricted MI to Kaggle data set due to its high memory and computation requirement on SHC data. For mean imputation we exploited the fact that most entries of each column are equal and therefore $X$ is sparse. MI leads to a large non-sparse matrix that significantly increases memory and speed requirements. Results have been omitted due to space constraints.

---

<table>
<thead>
<tr>
<th>Disease</th>
<th>Abbrev.</th>
<th>SHC</th>
<th>Kaggle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronary artery disease</td>
<td>CAD</td>
<td>75,619</td>
<td>1,096</td>
</tr>
<tr>
<td>Malignant cancer of any type</td>
<td>CANCER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>CHF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disorder</td>
<td>COPD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>DB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dementia</td>
<td>DEM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>PVD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Renal failure*</td>
<td>RF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe chronic liver disease*</td>
<td>SCLD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: List of diseases under study. Diseases marked with * were only studied on SHC data due to sparsity or missing labels in Kaggle.

<table>
<thead>
<tr>
<th></th>
<th>SHC</th>
<th>Kaggle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>75,619</td>
<td>1,096</td>
</tr>
<tr>
<td>Number of available lab exams</td>
<td>1,313</td>
<td>258</td>
</tr>
<tr>
<td>Median number of patients taking a lab</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>Median number of labs taken per patient</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>Minimum patient age</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Maximum patient age</td>
<td>110</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 2: Comparison of SHC and Kaggle dataset
competitive prices of various laboratory exam packages, which are comprised of several laboratory exams. These prices were used to determine the cost of the set of laboratory exams which were selected by our methods to be the most important for disease prediction.

We emphasize that most laboratory exams are not sold individually but in packages. Thus, in order to find the cost of our screening procedure, we solved a set-covering problem, formulated as a mixed-integer optimization problem where we minimized the cost over the possible sets of packages but enforced the choice of packages to include all of the laboratory exams sought. This problem can also be formulated as a convex linear programming problem, but since the size of the problem did not lead to a computational impasse, the mixed integer program was suitable for our needs. More specifically, we solved the optimization problem given by

$$\text{minimize} \sum_{S \in \mathcal{S}} c_S q_S, \quad \text{subject to} \quad \sum_{S \in \mathcal{S}} q_S \geq 1 \quad \forall e \in \mathcal{U}, \quad q_S \in \{0, 1\} \quad \forall S \in \mathcal{S}$$

where $\mathcal{S}$ is the set of all packages, $\mathcal{U}$ is the set of laboratory exams for which we wish to obtain the total minimum cost, $c_S$ is the cost of a package, and $q_S$ is an indicator variable which indicates whether a particular package $S \in \mathcal{S}$ is chosen for inclusion. We solved this optimization problem using the software of [24]. We then calculated the cost of the laboratory exams by adding the prices of the set of packages $S \in \mathcal{S}$ for which $q_S = 1$ at the solution of (1).

1.3 Statistical Methods

We denote the different disease prediction models in this section as follows. We will refer to our main model as the multitask learning model (MTL). For each value of a tuning parameter, $\lambda$, this model selects $M$ biomarkers as relevant to the disease prediction. We derive a model from this, the ordinary logistic regression model (OLR-M), which takes as input data the results of the $M$ biomarkers which have been selected by the multitask model at each value of the tuning parameter, $\lambda$. We compare both of these models, the multitask model and the OLR-M model, to the results of solving one individual disease prediction at a time, for each disease. We will henceforth refer to this model as single task learning (STL), since this type of model only involves a single disease prediction. First, we begin with logistic regression (LR), a preliminary for our models.

**Logistic Regression (LR).** We briefly describe logistic regression, which we will denote by LR, before describing our methods. Given $n$ $p$-dimensional data vectors, $X_i$, and labels $y_i \in \{0, 1\}$, $i = 1, \ldots, n$, logistic regression finds the optimal values of a $p$-dimensional parameter vector $\theta$ and a scalar $b$ such that the conditional probability of the observing $y_i$, given the data $X_i$ for all data vectors $i = 1, \ldots, n$, is maximized. This optimization problem is given by: $(\hat{\theta}, \hat{b}) = \arg \max_{\theta, b} \{ -\sum_{i=1}^{n} \log(P(y_i|X_i; \theta, b)) \}$ where $P(y_i|X_i; \theta, b)$ denotes the conditional probability that one observed the label $y_i$ given the data observation $X_i$, which is parameterized by $(\theta, b)$. We will refer to this term as the likelihood. This probability is commonly given the form $P(y_i|X_i; \theta, b) = 1/[1 + \exp(-y_i(\theta^\top X_i + b))]$. This form of $P(y_i|X_i; \theta, b)$ is used in this paper.

**Single Task Learning Model (STL).** The single task learning model formulates the goal of maximizing the log-likelihood while keeping the set of nonzero elements in the parameter vector $\theta$ to be small. This model consists of a logistic regression and a regularization term, commonly known as the LASSO penalty [25]. In the context of disease prediction, for each disease $j = 1, \ldots, K$, where $K$ denotes the total number of diseases considered, the model finds the set of parameters that maximizes the conditional probability that a patient will be diagnosed with disease given his or her vector of biomarkers. More specifically, for each disease $j$, we solve the problem

$$\left(\hat{\theta}_j, \hat{b}_j\right)(\lambda) = \arg \max_{\theta_j, b_j} \left[ -\sum_{i=1}^{n} \log \left( \frac{\exp(y_i^{(j)}(\theta_j^\top X_i + b_j))}{1 + \exp(\theta_j^\top X_i + b_j)} \right) + \lambda \sum_{m=1}^{p} |\theta_{jm}| \right]$$

where as defined previously, $n$ is the number of patients, $y_i^{(j)}$ is an indicator of whether patient $i$ had disease $j$, and $X_i$ is the vector of biomarkers for patient $i$, of length $p$. The vector $\theta_j$ and scalar $b_j$ are the learned parameters for the model for disease $j$, with $\theta_{jm}$ representing the $m^{th}$ element of $\theta_j$. If $\theta_{jm}$ is nonzero, then the model has selected this element as relevant for prediction. The magnitude of the parameter $\theta_{jm}$ represents the importance of biomarker $m$ to the prediction of disease $j$, and the larger this value the greater its importance for achieving higher
prediction accuracy. The first term in (2) is the logistic regression, and the second term is the regularization term with \( \lambda \) as the regularization parameter. The value of \( \lambda \) controls the effect of enforcing, for a particular value of \( j \), the number of values of \( m \) for which \( \theta_{jm} \) are nonzero to be small. A larger value of \( \lambda \) encourages the effect more strongly. Note that a different model is learned for each value of \( \lambda \).

**Multitask Learning Model (MTL).** The multi-task learning model formulates the goal of maximizing the log-likelihood and enforcing that the set of biomarkers which is jointly predictive of all diseases of interest to be small. This model consists of a logistic regression and a regularization term known as the GLASSO [15], or group lasso, penalty. In contrast to the STL model, we jointly learn the parameters for all disease predictions, by solving the optimization problem given by

\[
\left( \hat{\theta}_1, \ldots, \hat{\theta}_k, \hat{b}_1, \ldots, \hat{b}_k \right) (\lambda) = \arg \max_{\theta_1, \ldots, \theta_k, b_1, \ldots, b_k} \left[ -\sum_{j=1}^{K} \sum_{i=1}^{n} \log \left( \frac{\exp(\theta_{j}^T X_i + b_j)}{1 + \exp(\theta_{j}^T X_i + b_j)} \right) + \lambda \sum_{m=1}^{p} \left[ \sum_{j=1}^{K} \theta_{jm}^2 \right] \right]
\]

where all variables are as defined previously. As with the single task learning model, the first term is a logistic regression, and the second term is the group regularization term, which encourages biomarker \( m \) to be either nonzero or zero across all diseases together, i.e. that for each \( m \), \( \theta_{jm} \) should be nonzero or zero for all \( j \). Larger values of \( \lambda \) enforce this more strongly. Intuitively, such co-variation of weights through this group regularization can be interpreted as transfer of information between different disease predictions during the training of the model and can further lead to greater generalization ability of the model. When \( K = 1 \), the penalty term then reduces to the penalty term in the STL model, and the parameters are learned for prediction of only one disease; thus, this model reduces to the STL model. As with the STL model, a different model is learned for each value of \( \lambda \).

**OLR-M Model.** The OLR-M model is derived from the MTL model. After solving for the optimal parameters \((\theta_{j1}^*, b_{j1}^*)\), \( j = 1, \ldots, K \) for the MTL model with parameter \( \lambda \), we retrained a truncated model, as follows. For each value of \( \lambda \), we determined the nonzero biomarkers, i.e., the values of \( m \) for which the quantity \((\sum_{j=1}^{K} \theta_{jm}^*)^2\)^{1/2} was above a threshold. We will denote this number of nonzero biomarkers by \( M \), and give further details on how the number of nonzero biomarkers was determined, in the section Learning the Model Parameters. We then retrained the model over only these \( M \) biomarkers for each disease prediction, \( j = 1, \ldots, K \), formulated as an ordinary logistic regression.

**Learning the Model Parameters.** The optimization package MinFunc [26], which utilizes spectral projected gradient, was used for solving the associated optimization problems of the considered approaches, given for the STL, MTL, and OLG-M model described below. For each value of \( \lambda \), all models were trained and tested using five-fold cross validation. During training, due to the lack of positive examples, in each cross-validation fold (or bootstrap sample, in the case of the Kaggle dataset) the training data was sampled so that, for the SHC dataset, the ratio of positive to negative examples in the training set was 1:1, and for the Kaggle dataset it was 1:3. The main performance metric for all prediction tasks is area under ROC curve (AUC). All demonstrated AUC values in the figures are cross-validated, thus averaged, AUCs.

For each value of the regularization parameter \( \lambda \), we considered biomarker \( i \) to have been selected by the model if for any value of \( j \), it had a nonzero value of \( \theta_{ij} \). We determined this as follows. For the MTL model, for a particular value of \( \lambda \), the number of nonzero features was determined by counting the number of biomarkers \( m \) for which the quantity \((\sum_{j=1}^{K} \theta_{jm}^2)(\sum_{j=1}^{K} \sum_{m=1}^{p} \theta_{jm}^2)\) was greater than a threshold of \( \tau = 10^{-6} \) for at least one value of \( j = 1, \ldots, K \). For the STL model, we considered a biomarker as nonzero if for some value of \( j \in \{1, \ldots, K\} \), \((\theta_{jm}^2)/(\sum_{m=1}^{p} \theta_{jm}^2)\) was greater than \( \tau \). Thus each value of \( \lambda \) corresponds to a set of nonzero biomarkers, and a cost can be derived via solving the optimization problem (1), as given in the section Cost Data.

## 2 Results

### 2.1 Biomarker Selection for Multiple Disease Prediction

We compare the results of the MTL, OLR-M, and STL models for disease prediction, number of biomarkers necessary to achieve the given results for prediction, and cost. As our evaluation metric, we will use area under
the receiver-operating characteristic (ROC) curve, henceforth abbreviated as AUC. The AUC values shown in the succeeding figures are averaged over the five cross-validation folds, as well as averaged over the nine different diseases of interest.

The MTL and OLR-M models achieve comparable accuracy to the STL model with a smaller number of biomarkers. Figure 1a presents this result visually, in which the given curves show the effect on the value of the AUC is when the number of biomarkers selected by the model is varied, for each of the different modeling approaches. This results also translates when considering the cost of the biomarkers needed to achieve each AUC value on the curve. Figure 1b presents the same result as Figure 1a, but calculates the cost of the biomarkers at each point on the curve in Figure 1b. Thus the effect is amplified when comparing cost of the MTL and OLR-M models to the STL model, as the MTL and OLR-M models are much lower in cost. The OLR-M was the most succinct of all the considered approaches, and in particular, the value $M = 30$ was found to be the optimal value for $M$ using cross validation to have the highest AUC, as presented in Figures 1a and 1b.

![Figure 1: Comparison of number of biomarkers selected and associated cost by the MTL, STL, and OLR-M models. Each point on each curve is obtained by changing the regularization parameter.](image)

2.2 Sensitivity Analysis

We study the effect of varying the length of, as well as a shift in, the time interval over which we identify a patient as having a positive diagnosis. In the previously presented results, the time horizon considered for diagnosis was one to thirteen months after biomarkers were taken. We consider a shift of a year, i.e. a time horizon of twelve to twenty-four months after the biomarkers were taken. We also study the performance when the time horizon considered is varied, one to six months, and one to eighteen months, after the biomarkers were taken. Due to space limitations, in Figure 2 we only demonstrate the results for twelve to twenty-four months interval and one to six months interval. In all cases, we observe the same pattern as in previous results.

We also compare predictive accuracy of the three models using other metrics than AUC. In particular, Figure 3 shows that the three models have comparable accuracy with respect to sensitivity (or true positive rate). However, with respect to positive predictive value, the MTL and OLR-M models are indistinguishable and closely follow the STL model.
(a) Cost comparison, diagnosis one to two years after lab results
(b) Cost comparison, diagnosis during a six month time horizon

Figure 2: Comparison of costs and accuracy of models, for a shift and varied lengths of time horizon.

(a) Average sensitivity versus fraction of selected patients
(b) Average positive predictive value versus fraction of selected patients

Figure 3: Comparison of predictive accuracy of models based on sensitivity and positive predictive value.

2.3 Biomarkers Selected for OLR-M Model

As aforementioned, the value of $M$ which achieved the highest AUC, when cross-validated, was $M = 30$. We have given a list of the top $M = 30$ labs selected by the OLR-30 model in Table 3. All of these labs are associated with the diseases of interest in this paper, and as expected from the design of our method, many of them are associated with several diseases. For example, abnormal glucose levels are associated with signs of diabetes, cancer, renal disease, liver disease, and heart disease. Similarly, abnormal alkaline phosphatase levels can be associated with liver disease, cancer, as well as heart disease. A vitamin B12 test can be given to those experiencing symptoms of anemia, as well as dementia and memory problems, but higher levels can also be associated with diabetes, liver disease, and heart problems. Additionally, the considered model also has the ability to discover associations between these biomarkers and the diseases which may not be explicitly known; hence, the associations go beyond known ones, some of which have just been mentioned.
Table 3: List of the top features with their rank determined by OLR-30. Smaller rank means the feature group received a larger value of $\sqrt{\sum_{j=1}^{K} \frac{\hat{\theta}_j^2}{\hat{\theta}_j^2}}$ in multi-task learning.

2.4 Generalizability of OLR-M to Different Patient Populations

We investigate the performance of the OLR-M model on a different patient population, the Kaggle dataset. As aforementioned, the Kaggle dataset is a less ideal dataset, with fewer biomarker values available per patient. As $M = 30$ achieved the best predictive accuracy and lowest cost when validating on the SHC dataset, we will use this model, and refer to it as OLR-30. The OLR-M model is able to achieve comparable accuracy to the STL model, with far fewer biomarkers, and thus has similar performance on this dataset as on the SHC dataset. The OLR-30 model consists of 30 biomarkers, as aforementioned, much fewer than the STL model, which requires the results of 63 biomarkers, in addition to age and gender information. Table 4 provides the confidence intervals at the 5% level for the AUC values for both models, for each disease under study. The results were obtained using 5-fold cross validation and the method of DeLong [27] for the confidence intervals.

Remark. The Kaggle dataset, as aforementioned, is a lower quality dataset than SHC, which required us to process the data differently than the SHC dataset as discussed in the section Patient Data. However, we observe that the results follow the same pattern as the results for the SHC dataset, as presented in Table 4.

Table 4: Confidence intervals for the AUC values of OLR-30 and STL approaches

<table>
<thead>
<tr>
<th>Disease</th>
<th>OLR-30</th>
<th>STL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD</td>
<td>0.7535 ± 0.0468</td>
<td>0.7412 ± 0.0483</td>
</tr>
<tr>
<td>CHF</td>
<td>0.6003 ± 0.0489</td>
<td>0.5961 ± 0.0476</td>
</tr>
<tr>
<td>COPD</td>
<td>0.6492 ± 0.1067</td>
<td>0.6604 ± 0.1043</td>
</tr>
<tr>
<td>DEM</td>
<td>0.6359 ± 0.1192</td>
<td>0.6447 ± 0.1131</td>
</tr>
<tr>
<td>PVD</td>
<td>0.9184 ± 0.0716</td>
<td>0.9183 ± 0.0676</td>
</tr>
<tr>
<td>RF</td>
<td>0.7957 ± 0.0684</td>
<td>0.7925 ± 0.0656</td>
</tr>
</tbody>
</table>

3 Discussion

We have proposed two approaches, the MTL and OLR-M models, for developing a clinical model for multiple disease prediction. Our study has focused, in particular, on how to design a health-risk assessment (HRA) for prediction of multiple diseases, using statistical methods and patient data. In contrast to the STL model, which is one appealing and commonly used statistical approach in the literature, we propose a solution which relies on
group regularization methods to jointly learn a small set of biomarkers are the most relevant for prediction, for multiple diseases.

Although our results are encouraging, we note that our study has limitations. Although we would like to also compare the prediction accuracy of our method against common scoring rules, such as the Framingham score, used for individual diseases, we are unable to do so at this time because of the unavailability of necessary biomarkers in our data for these scores, such as blood pressure, whether the patient is a smoker, or whether they are on dialysis. Hence, we limit our comparison to the state-of-the-art statistical method for disease prediction, STL. The lack of temporal information regarding disease diagnosis in the Kaggle dataset is surely a limitation. Validation on this dataset has allowed us to compare the OLR-M and STL methods, however, we cautiously use these results to guide our conclusion that the OLR-M model is a viable option for a low cost multiple disease prediction model.

The results of our investigation show that the MTL and OLR-M models provide one solution to designing a lower cost HRA. We have presented experiments to validate that our models achieve comparable accuracy to the STL model, but achieve this accuracy with the use of fewer biomarkers, which directly translates to cost. Our MTL model uses group regularization to select a small set of biomarkers, which can be used to jointly predict all of the diseases at once. By learning which biomarkers are jointly successful for prediction, the MTL model is able to achieve comparable accuracy to the STL model, but with fewer biomarkers. The OLR-M model enhances these benefits by producing a model with fewer biomarkers than either of the other two considered approaches, and keeping a high accuracy by refining over the MTL model. Since the MTL model is a biased model, the OLR-M acts to remove the bias from this model, increasing accuracy of the MTL model at the points for which it has the highest AUCs.

The same pattern of results is observed when we consider different time horizons for considering a positive diagnosis of a patient, i.e, that the MTL and OLR-M models produce lower cost models with comparable accuracy. We have studied the performance of the approaches under a one to six month time horizon, as well as twelve to twenty-four months, when there is a lag of a year between when the patients’ biomarkers were taken and when the positive diagnosis occurred. We observed similar accuracy between the models as well when we studied the performance under other metrics, other than AUC.

Lastly, we have validated the OLR-M model on a new patient population, that of the Kaggle dataset. This dataset is a significantly smaller dataset, with fewer patients, fewer biomarkers per patient. However, we observe the same result to hold for this new dataset, as the OLR-M model is still as accurate as the STL model but much less costly. Thus we believe the the performance of the OLR-30 model is not restricted to the SHC dataset from which it was derived, but can be extended to new populations.

References


Physician handoffs: opportunities and limitations for supportive technologies

Katherine S. Blondon, MD, PhD1,2, Rolf Wipfli, PhD1, Mathieu R. Nendaz, MD, MHPE1,2, Christian Lovis, MD, MPH1,2
1University Hospitals of Geneva, Switzerland; 2University of Geneva, Switzerland

Abstract

Shift-to-shift handoffs refer to the process of transferring role and responsibility for providing care from one person to another, thus insuring continuity of care. Through focus groups of residents and supervising physicians, we studied how physicians select patient cases to discuss during handoffs. We also compared the selection across level of experience. Understanding the patient selection criteria can give us insight into how to improve handoffs, in particular using supportive technologies that are integrated into the clinical information system. Studying the actual handoff process and note-taking also generated suggestions for handoff improvement.

1. Introduction

Transitions of care can endanger the quality of patient care and patient safety. Care transitions occur between physicians during patient transfers from a ward to another, or during shift changes. The handoff process involves the transfer of role and responsibility from one person to another1. Communication modalities range from in-person encounters to phone calls or written sign-outs. Studies of adverse events established that up to 60% of preventable adverse events are related to communication failures2. Improving handoffs can be challenging, as the modalities and processes vary widely from one department to another.

This study focuses on the evening handoffs between usual day teams (Monday to Friday) and the night teams (who also cover Saturdays and Sundays). During this evening handoff, the dayshift teams discuss a selection of patients from their wards with the nightshift team in presence of the supervising physician, covering for the weekend. Not all patients will be discussed. Patients thought to be stable, for example, are usually not presented during evening handoffs. Weekend shifts work on the same model, with an additional written handoff for all patients on Friday evenings. While prior studies on handoffs have suggested ways to standardize the overall content of handoffs with the use of checklists3, for example, or to structure the presentation of each case with mnemonics4, little evidence supports the process underlying the choice of patient cases to be discussed.

Handoffs are complex procedures, requiring the integration of a large amount of information about each patient, and errors can lead to high costs. Job aids were designed to “extend human capability to store and process information5,” and have proved to be especially useful for tasks with these characteristics6. Yet, the design of useful job aids, which can help decrease the cognitive load, require a clear understanding of the tasks and reasoning procedures. As stated in the Cognitive Load Theory, “An element is anything that needs to be understood and/or learned. If elements interact, they cannot be understood in isolation7.” The multiple interactions between pertinent features contribute to the complexity of handoffs, and should influence how the job aid is designed.

Selective handoff processes, where only certain patients are handed off, emphasize the need for appropriate patient selection. Understanding this selection process can have implications both for educational interventions about the handoff preparation process, as well as for the design of supportive technologies. The current electronic medical records include many features, such as admission, progress and discharge notes, labs, radiology images and reports, reports of all other ancillary tests, nursing documentation, and a dashboard for the healthcare users. It does not, however, have a specific handoff tool. Notifications and alerts can be potentially used to help identify patients for handoffs, but the criteria for such notifications need to be well chosen, to avoid forgetting to hand off a patient, and to avoid alert fatigue8. The goals of this study are to identify patient selection criteria, and to compare the approaches of physician residents and supervising physicians. We present a framework for patient selection, which was similar across level of experience, and suggest ways

2. Methods

2.1 Data collection

After approval by the State ethical review board, we conducted four semi-structured focus groups of five to seven participants between February and May 2014: two with residents, and two with supervising physicians. We recruited participants from the Division of General Internal Medicine through emails and general announcements (i.e., prior to
grand rounds). We excluded the senior attendings, as they do not take part in the evening handoff process. Using eight standardized clinical vignettes of typical internal medicine patients, participants started by individually selecting the patients that they would hand off for both a weekday and weekend handoff session. The cases are detailed in Table 1. The choices were reported on a board to facilitate the group discussion. Then as a group, participants shared the reasons of their choices, and reached a consensus handoff. We completed the session with an open discussion about more general questions on personal experiences, in particular during their on-call months, during which they were on night and weekend shifts. We also enquired about the importance of the order in which cases were handed off, and whether handing off to a supervising physician (rather than a resident or medical student) would have affected the content. We asked about the use of mnemonics for the verbal handoff. All sessions were conducted in French.

Table 1. Overview of standardized cases

<table>
<thead>
<tr>
<th>Cases</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>Prior nocturnal agitation in a 79 yo male, recovering from spleen rupture, pneumonia and renal failure, and who fell from his bed two nights ago</td>
</tr>
<tr>
<td>Case 2</td>
<td>Fever and hypotension in a 79 yo female with diabetes</td>
</tr>
<tr>
<td>Case 3</td>
<td>Subacute onset of fever, in a context of bloody diarrhea in a 38 yo male</td>
</tr>
<tr>
<td>Case 4</td>
<td>Asymptomatic INR of 6.2 (excessive anticoagulation) in a 84 yo female with diabetes, atrial fibrillation and currently admitted for pneumonia</td>
</tr>
<tr>
<td>Case 5</td>
<td>Malaise at time of handoff in a 71 yo female with ischemic heart disease, atrial fibrillation and diabetes, admitted for weight loss and anemia</td>
</tr>
<tr>
<td>Case 6</td>
<td>Abdominal pain during chemotherapy in a 67 yo male with non-small cell lung cancer</td>
</tr>
<tr>
<td>Case 7</td>
<td>Iliac pain &lt;24 hours after medullary biopsy during work-up for multiple myeloma for in a 66 yo male with chronic lumbar pain and depression</td>
</tr>
<tr>
<td>Case 8</td>
<td>Shortness of breath in a 59 yo male with emphysema, admitted for suspected lung cancer</td>
</tr>
</tbody>
</table>

2.2 Study setting

Medical teams in the Department of General Internal Medicine at the largest Swiss teaching hospitals organization are responsible for patients throughout the day from 8 am to 6:30 pm during the week. At 6 pm, the day teams hand off their patients to the nightshift resident. The Department is split into two zones, each of which is managed by one nightshift resident until the next morning. Each nightshift resident cross-covers about 80-90 patients from five different wards, which means they are responsible for all of these patients, regardless of whether they are familiar with their cases or not. Although nightshift residents have in-house supervision, the supervising physician actually works in the Emergency department, and usually only discusses urgent or complex cases with the nightshift residents.

Between Mondays and Thursdays, day teams typically leave written progress notes as needed, i.e. for a procedure performed during the day like a pleural tap, or for a new management approach. On Fridays, however, all day-teams are required to leave a full progress note for each patient, which also serves as a written handoff, since it also includes which parameters to monitor, tests results to track, and a general list of problems.

2.3 Data analysis

All discussions from the focus groups were transcribed and de-identified for subsequent coding with thematic content analysis. Two of the investigators coded the transcripts, iteratively comparing and contrasting themes
between individuals, cases and focus group sessions. We also analyzed according to the level of experience (resident vs supervising physician). The results of individual patient selection from the board were also analyzed using descriptive analyses, and a multi-rater kappa analysis by level of experience.

3. Results

3.1 Participant characteristics

We enrolled 12 residents and 10 supervising physicians. Residents had an average of 2.8 years of postgraduate clinical experience (range of 1 to 5 years), and supervisors had an average of 6.8 years of experience (range of 5 to 9 years), with 2.4 years on average of supervision. These levels of experience are representative of the current physicians in our Department (senior attendings not included in this study), with board certification occurring on average after 5 years experience. There were 7 female supervisors, and nine female residents. Six supervisors and three residents reported some prior handoff training. All participants had completed at least one 3-month rotation in emergency medicine.

3.2 Patient selection

Our study focused mainly on patient selection for verbal handoffs. Participants identified several criteria to help identify patients that needed to be handed off to the night-shift physician or weekend teams. Despite some degree of overlap, we can categorize the criteria. A) related to patient safety, such as unusual or complex cases, or information that might be difficult to find in the EMR. B) related to care workflow, such as tracking tests results, or clinical monitoring. C) related to acute situations, with recent aggravation or unstable patient that need close monitoring. D) patient comfort, and end of life care. Finally, E) related to the prevention of complications or F) events that are anticipated. These themes are illustrated in Figure 1.

![Figure 1. Criteria for patient selection.](image)
Patient safety

Patient safety involves issues in clinical reasoning and information retrieval. Participants handed off cases with atypical management, to avoid confusion or even errors by the night shift physician. Supervisor SP3 explains: “We had to choose an unusual antibiotic for one of my patients, a choice that could have been questioned by the night physician... Sometimes, when we explain why we decided to choose B rather than A, the night-shift physician won’t waste time wondering why the day-team chose this or didn’t do that. So I like to handoff unusual approaches.” SP5 also gave an example of this: “For patients with fever of unknown origin, we want to avoid antibiotics. The spinal reflex for fever usually is to administer an antibiotic. So I like to hand off when to give which antibiotic, and to explain why we choose to NOT give an antibiotic to someone with fever.”

Cross-covering physicians are not familiar with the patients they are responsible for, and are often under time-pressure. Day-time physicians are much more familiar with the patients’ conditions and have time to discuss patient management with their supervising physician and consultants. Providing a brief explanation for certain diagnoses can help the night physician make decisions, or at least help save time when reviewing the medical charts. One resident (R11) said about a patient with chronic hypokalemia from renal loss: “The day-physician who has been managing this patient over the last week knows her situation by heart, particularly for potassium substitution, whereas the night doc may not have the time to review the case in depth.” R3 describes her own experience during night-shifts: “It’s not because we don’t want to think, it’s just to help understand where the problem lies because nights are busy, we are not familiar with the patients, and sometimes it’s just tricky to go over the whole story when it’s urgent.”

Another reason to hand off a patient was if they thought that some medical information might be difficult to find in the EMR. For example, SP2 reported an example with dialysis: “For dialysis, the information needs to be more detailed. Urgent dialysis takes place in the ICU. So for patients – I remember this patient who had a “Do not resuscitate” order, yet would have been admitted to the ICU for urgent dialysis. It wasn’t well handed off, and that information was hard to find...” The goal then of handoff in this case would be to raise awareness about certain aspects of patient management.

Care workflow

Participants also handed off patients who had pending imaging, on-going labs or other tests, so that the on-coming physician could track results in a timely manner, and adapt a patient’s management if needed. SP5 discusses tracking as a handoff goal: “[...] tracking exams that have been ordered, so there’s no chest x-ray with a white-out that is missed until the following morning. So pending lab tests and imaging should be tracked by the night-shift to avoid late discoveries.”

Acute care

Patients who were considered unstable, or who had a recent aggravation of their clinical state, were also handed off. R9 explained: “Something acute is happening, and we are starting initial measures. But we need to see how the patient will respond to all this, so we can’t assess if the patient is stable or not at this time.” Handoffs allow patients to be seen rapidly for an initial assessment, and then to be monitored during the night, ideally with clear recommendations for pain, blood pressure or diuresis, for example.

Receiving a handoff about an unstable patient also helps the on-coming physician to attribute priorities to the numerous tasks she has to accomplish. The day team sometimes emphasizes the need to see a patient soon to reassess the situation. As stated by one supervising physician (SP5): “One of the goals of handoffs is to help the night resident decide the order in which to do things. By handing off [an unstable patient], she might start by assess this patient, rather than just come across her later during her shift, and he might realize that the diuresis decreased over the past two hours. If instead she does receive a handoff about this sick patient, she might go see her earlier.” Nursing concerns can also contribute to a physician to hand off a patient. Nursing teams are more present at the bedside than physicians, and may be more able to assess changes in status. Nurses are also often the first person to detect or suspect a cause for concern in hospitalized patients. In the case with suspected lung cancer, SP4 explains: “I would hand off this patient because the nurse assessed him to have increased shortness of breath. It’s a vague symptom, and because it’s vague, I will probably need to hand him off for more investigation. I believe the day resident should go and evaluate the patient, to gather more information, so that the night physician can reassess the patient later in the evening. There may also be a chest X-ray to track.” In this example, we see that there are often several reasons to hand off a patient, first due to a yet unexplained aggravation in the clinical status, which then needs clinical monitoring by the night physician and potential test result tracking.
Prevention

In order to maintain continuity of care for patients, participants emphasized the need to anticipate potential problems during the night or weekend. Anticipation requires good knowledge of a patient’s case, where recurrence of an event is possible, or with good knowledge of a clinical situation, such as a common complication of a procedure. An example of a recurrent event, is described by a supervising physician (SP2): “[... because this patient had a high probability of presenting another episode of agitation, he’d had them often in the past.” Another supervising physician anticipates a complication of a recent intervention: “I would discuss this patient simply because he had an intervention the day before, and because he’s still in the time frame for a complication. It’s not just the pain, it’s also about seeing if there’s bleeding or a hematoma” (SP4)

Patient comfort

Participants also agreed that patients with end of life care should be handed off, so the night shift physician could (1) assess patient comfort and adapt treatments if needed, (2) not be alarmed if the nurse calls about the time of death, and (3) be aware who to inform about changes in condition or death (and in what delay this needed to be done). As illustrated by SP10’s statement: “It’s to let us know that if the clinical state worsens, to provide end of life care rather than aggressively reanimate. It’s also to inform about whether the family wishes to be called any time of the night, to know where to find their phone number…” (SP10)

Other reasons, divergent opinions and use of written handoffs

Other reasons to handoff a patient were to inform about patients with multiple or recurrent symptoms of unclear origin (after excluding the treatable or urgent etiologies), such as “chest pain, which has already been worked up several times, cases where clinicians tend to worry and begin the work-up all over again each time the symptom appears” (R9).

Physicians also reported relying on nursing teams for part of patient monitoring. In some situations with divergent opinions about handoff, some clinicians advocated mainly relying on the nursing team, providing clear targets for monitoring, and expecting nursing teams to call the night physician if the targets were not reached. SP1 described this for the case with sepsis: “[…] an acute event in a hospitalized patient […] after a clinical assessment, I would decide about prescribing an antibiotic and close nursing surveillance for blood pressure and diuresis. I think the nursing team has a role to play here, and it’s the nurses who will call the doctor if needed. So for the night physician, I don’t see what more the doc could do if just checking in spontaneously, if the monitoring targets are well established.” For other participants, handing off such a patient was a potential a teaching moment for the night resident, about what to do if the targets were not achieved. Handoffs in this case were also about deciding when to call the ICU to transfer the patient.

In several of the discordant opinions about handoffs, the debate was not so much about whether to hand off a patient, but more about whether the handoff should be verbal or written. Oftentimes, simple to-dos such as weight monitoring or labs to track are not handed off verbally, but will be highlighted in the written handoff notes for the weekend day physicians.

3.3 Resident vs supervising physician approach

Overall, there was clear agreement on the ultimate goal of handoffs, which is to maintain continuity of safe and effective healthcare. Participants also clearly stated differences in patient management during day and night shifts, where night teams manage the acute events, and follow recommendations made by the day teams. During the day, patient management includes a more in-depth study of patient cases, addressing more chronic issues, work-up, and discussions with consultants.

Using descriptive statistics, our results show similar proportions of evening handoffs and weekend handoffs between residents and supervising physicians. (Figure 2)
Figure 2. Proportion of cases handed off by level of experience. There was no significant difference between cases selected by residents or by supervisors.

The multi-rater kappa statistic showed moderate agreement for patient selection among residents and supervisors for both weekday evening handoffs (0.53 and 0.65 respectively) and for weekend handoffs (0.44 and 0.58).

Handoff rates were lower during the weekends. In our usual practice for weekends, day-teams prepare a written sign-out for the weekend teams, with a summary of current problems, and a list of to-dos, with the context (relevant past history). A complete sign-out is often not prepared for weekday handoffs, and only a brief note is left when a procedure is performed or if the patient’s condition changes.

3.4 General comments about handoffs

There was also an overall agreement about the content of a verbal handoff. They were expected to be concise, with clear day-team recommendations for anticipated problems (if-then statements) and a to-do list of tasks, such as tracking a lab or other ancillary test results, or clinical monitoring. Participants all felt that keeping handoffs brief helped to avoid “diluting” the information.

While there is no formal handoff training in our institution, clerks and residents receive a pocket card with a mnemonic for handoffs when they start their employment in our department. This mnemonic suggests a way to present the verbal handoff, but does not concern the patient selection process. Participants did not use this pocket card or mnemonic, some barely remembered having this card, because they considered that their current process was close to the recommendations in the mnemonic (“I read the card, and it seemed to resemble what I currently do, so…” R2). The card seemed to have more use as support when explaining handoffs to the medical students. Day teams sometimes discuss the patient selection with the supervising physician. The handoff itself can be given by a resident, a final-year clerk, or more rarely by a supervising physician.

The order in which patients were presented did not matter to the participants, as they mainly needed to know the ward in which the patients were. Participants all confirmed the importance of interactions during the handoff, for clarification about recommendations, general management decisions, or simply for precision.

Participants mentioned the importance of taking notes while receiving handoffs, as a way to remember key features for a patient. The current practice for note-taking in our practice is to start with blank paper. Participants usually ask the day teams to give the ward name before beginning their handoff, in order to take notes at the right place in their sheets. In most cases, the handoff begins with the name of the patient, year or date of birth, resuscitation order, followed by medical information and “to-do list”. Although the spelling of patient names may be incorrect, and despite potentially missed items while annotating, participants were quite satisfied with the current process. There is no access to the electronic medical record (EMR) during the handoff, but participants reported using the EMR after the handoff process to clarify the information received, if uncertainties remained after the handoff process. Both residents and supervising physicians mentioned that improvements in handoffs should not lead to an increase in workload for the day-teams, particularly since many already work overtime on Fridays preparing written handoffs for their patients.

Last but not least, participants agreed that knowing who they are handing off to can change the content of the handoff. Clinicians tended to provide more detailed information to junior physicians, with more recommendations.
4. Discussion and design implications

Understanding the residents’ and supervisors’ perceptions and expectations of the handoff process is the first step to improving handoffs. Participants agreed on the main goal of handoffs, which is to maintain the continuity of care between day and night teams through a transfer of responsibility. Therefore, there was overall agreement over the need to (1) identify patients at risk of possible serious events in the near future, (2) explain unusual or complicated reasoning processes, (3) identify patients with pending tests, and (4) provide recommendations for the night or weekend teams, in particular with explicit if/then statements.

Improving handoffs can be achieved through combined approaches. We propose ways for supporting handoffs in Figure 3.

Figure 3. Approaches for improving handoffs. Dark blue boxes describe the sequential steps of handoffs, whereas the white boxes suggest approaches to support each step.

Support for patient selection:

Supportive technologies for handoffs can address some of the patient selection criteria used to identify patients who need to be handed off. Despite an overall agreement about these criteria based on our standardized clinical cases, many criteria are based on clinical reasoning processes, standard processes in patient management and on the expected evolution over time. Providing technological support for these criteria may therefore be more difficult than for other criteria such as pending tests.

Simple notifications or alerts could be created for patients who await some kind of imaging or lab test, or who have had a test with a pending result. If these notifications are visible in an overview mode, they could help the day-team identify patients to hand off for result tracking and subsequent management. Designing such a system could have additional benefits for certain time-consuming tests, such as cultures for mycobacteria, which can take several weeks or months. Currently, some of these results may be missed altogether; or they may be seen by the hospital physician but not the ambulatory physician if the patient has been discharged. Receiving some form of notification for these tests could help improve the communication process.

Anticipation seems to be a key component for patient selection for handoffs, as it is required when assessing risk of problems occurring in the future. Anticipation requires an understanding of the patient’s diagnoses, and of how they will evolve over time. Clinical evolution of a given diagnosis has been studied in the form of clinical pathways, or trajectories\(^\text{10}\). We could imagine having a supportive system that could notify the physician when a patient’s evolution diverges from a clinical pathway. For example, for patients with pneumonia, fever should abate within 72 hours of the start of antibiotics. If this is not the case, the supportive system could notify the day physician as she prepares for handoffs at the end of the day. Ideally, not only vital signs, but also lab results should be integrated in the expected trajectory. The difficulty, however, is that the effect of a patient’s comorbidities and age on the time-frame of the trajectory is less well studied. Furthermore, it is not just the absolute value that is of importance for a trajectory, but its relative value, particularly for changes in settings with chronic diseases (i.e., creatinine values with
and without chronic kidney disease). This trajectory-based approach has been discussed also for patient prioritization\(^\text{11}\), when a physician chooses the order in which to perform daily tasks for a set of patients.

Another advantage of this trajectory approach would be its ability to detect early warning signs for a patient starting to be unwell. Although nurses take the vitals during the nursing rounds, changes in status may not be immediately apparent, or may initially go unnoticed if the change is small. There may then be a delay until the doctor sees the change, or until the nurse discusses her concern with a doctor. A trajectory-based approach could also serve as an early warning system for patient deterioration.

One of the main challenges about notifications and alerts is the target physician and the threshold to activate the alert. We propose that for handoff support, day physicians should see a notification in an overview mode of the ward. The cause of the notification (pending tests vs divergent trajectory) should be visible, with easy access to the pertinent part of the chart for clarification. Another challenge about notifications is the threshold of abnormality\(^\text{12}\), particularly for easily fluctuating variables such as pulse or blood pressure.

Notification systems about pending results or divergent trajectories would benefit greatly from a global view of the patients in a ward. For the nightshift team, this could even be for all 4 or 5 wards that they cross-cover. The current EMR system at our institution has two overview modes for a ward: the first mode is a table, which lists all patients in a ward in an alphabetical order, with information on gender, date of birth, and date of hospitalization. The second mode is a synoptic view with the rooms and emplacement of patients within the ward. It also contains information about patient flow, with lists for the in-coming and out-going patients. The nursing version of the synoptic view already has information about whether new orders have been placed and whether they have been carried out by the nursing staff. This synoptic view could also present information about pending tests or pending results for each patient.

The form of notification is also important, in particular to avoid information overload\(^\text{13}\), or alert fatigue\(^\text{14}\). We propose a design with more subtle modified background for a trajectory-based approach, with hues according to divergence with the planned trajectory. Thus, the system informs the physician about a degree of “sickness” or of need for attention without giving explicit notifications or recommendations. Humans interpret a value quicker in a graphic form than in a text form, such as alert messages\(^\text{15}\). Furthermore, this graphical representation reduces cognitive workload\(^\text{15}\). The visual representation may further point to more detailed, textual information. For the pending test results, we propose a notification on the overview pages only, as our EMR summary cockpit page presents the most recent results, with an agenda for future tests.

Detection of unusual or complicated patient cases is difficult to support through technology. Our system currently does not track the number of consultants involved in a patient case, but this could be a proxy for the complexity of a case. The trajectory-approach, if detailed enough to identify therapy choices, could possibly be a way to help detect unusual patient management strategies. The choice of antibiotics, for example, could help differentiate a standard management from an usual process (e.g., resistant bacteria). On the other hand, physicians can usually identify unusual management choices, and may not need supportive technologies for this selection criterion.

**Support for handoffs**

Although observing the actual handoff itself was not the main goal of our study, participants all did mention taking notes when receiving the handoff. The preparation for the handoff could also include a written component, which could help improve the note-taking process. While we recognize the desire and need to make personal annotations about each patient, according to individual mental models, we also see potential areas for improvement. We could imagine using a print-out with key words, patient identity and basic demographics for the selected patients, for example.

As our hospital prepares mobile versions of our current EMR, we can suggest a design of brief patient summaries, easily extractible from progress notes or problem lists for each patient. By providing the patient name, and main problems, we hope to help the receiving physician focus on the handoff. These key problem lists should be expandable, thus providing information about the evolution, ancillary tests and consultations from the medical specialties when required. We do however need to avoid overwhelming the receiving physician with the written information, which (1) is a distraction from the verbal handoff, and (2) should not require additional paperwork for the day teams. Designing for mobile interfaces, due to their limited screen size, requires mindful selection of the pertinent information. The physician may access the deeper levels of information (i.e., specific laboratory results, more detailed information about on-going work-ups) when required by using the touch screen. Mobile devices may thus respond well to the needs of physicians during handoffs.
4.1 Limitations

The generalizability of our findings is somewhat limited by differences between our local practices and those of other institutions, particularly in terms of work organization and local EMR system. This variability in handoff procedures is a barrier to handoff improvement, and has motivated recommendations for standardization. Although some EMR systems may already have parts of the features suggested here, we have presented design implications that are based on our local EMR system, with which our physicians are familiar. We also note the use of fictitious cases, based on short vignettes, which differ from real patient cases that physicians manage. Although focus groups can lead to biased discussions due to peer pressure, we included an individual selection phase prior to the group discussion. We had a slightly higher representation of female physicians in our sample compared to our actual population, in both resident and supervising physician groups, which may affect some of our results.

Among the strengths of this paper, we note the local practice of patient selection for handoffs, which emphasizes the anticipation of potential problems in the patients near future, and allows the clinicians to keep their focus on the sicker patients. Even if this patient selection process may not be found in other healthcare models, our findings can help guide improvement in the handoff content, by focusing particularly on the features we identified for patient selection. We did reach saturation with our four focus groups. Furthermore, the use of standardized cases for the discussion allows for a comparison across focus groups and across level of experience. Participants found the cases realistic, and could relate to similar situations in their own experience.

5. Conclusion

Our focus group sessions with residents and supervising physicians used standardized patient vignettes to help study how physicians assess and anticipate needs in patient management. Patients considered to be unstable, or with yet unclear evolution after an acute event, those with unusual management strategies, those with complex medical cases, those who need monitoring or follow-up on tests, and end of life situations are selected for discussion during handoffs. Better awareness of these situations could help facilitate the handoff preparation process.

Future works

Based on our findings in this study, we plan to conduct usability studies with our new design implications for supportive technologies for handoff preparation. Using a mock-up of our current EMR, we plan to assess physician impressions and usability with the suggested changes, in particular for the synoptic view of an entire ward with notifications. Implementation of the notifications or alerts suggested above will require not only changes in the EMR system, but also further reflection how to reach the appropriate target, and will take much longer for implementation.

After studying the handoff preparation process, we are currently conducting a study on the integration of verbal handoff information, with the subsequent chart-biopsy process during cross-covering, followed by the participants giving their own handoff. We are comparing two interventions with the current note-taking (control): EMR access during the verbal handoff and use of a paper summary as suggested above (including patient identification and a list of problems). Studying the participants’ handoffs in this new study will provide further insight into the handoff preparation process.

References


Model Checking for Verification of Interactive Health IT Systems

Keith A. Butler, PhD¹, Eric Mercer, PhD², Ali Bahrami, PhD³, Cui Tao, PhD⁴
¹University of Washington, Seattle, WA; ²Brigham Young University, Provo, UT, ³Medico Systems, Bellevue, WA, ⁴University of Texas, Houston, TX

Abstract
Rigorous methods for design and verification of health IT systems have lagged far behind their proliferation. The inherent technical complexity of healthcare, combined with the added complexity of health information technology makes their resulting behavior unpredictable and introduces serious risk. We propose to mitigate this risk by formalizing the relationship between HIT and the conceptual work that increasingly typifies modern care. We introduce new techniques for modeling clinical workflows and the conceptual products within them that allow established, powerful modeling checking technology to be applied to interactive health IT systems. The new capability can evaluate the workflows of a new HIT system performed by clinicians and computers to improve safety and reliability. We demonstrate the method on a patient contact system to demonstrate model checking is effective for interactive systems and that much of it can be automated.

Introduction
The importance of interactive computer systems in healthcare has grown rapidly with the proliferation of networks of personal devices, laptops, and desktops. Rigorous methods for design and verification, however, have lagged far behind. Standard methods, such as usability testing, that were adequate for interactive systems used in business offices and homes, were never intended for safety-critical domains 15, 16. The inherent technical complexity of healthcare, combined with the added complexity of health information technology (HIT) makes their resulting behavior unpredictable and introduces serious risks 1. These risks threaten patient safety, disrupt care, undermine the added value of HIT, and impede its adoption.

Model checking is one technology that has great potential to help mitigate the risks of unpredictability in complex safety-critical systems 12, 13. A model checker proves whether a subject system satisfies the properties that it needs to meet in every possible behavior of the system. Applying model checking directly to interactive systems that are employed by clinicians, however, presents distinct challenges. One of the most severe is the difficulty to define clearly the conceptual work, such as disease diagnoses, treatment plans, surveillance, case-management, hospital admissions, and schedules for scarce equipment that is common in modern healthcare. Clinical workflow studies have shown these types of conceptual work products to be fundamental in HIT and care.

Until recently the problem of how to specify conceptual work products in a sufficiently rigorous manner prevented model checking of interactive systems for this aspect of care. Theoretical advances in cognitive science 17 combined with standards for declarative knowledge modeling 7, 21 now offer a technique to clearly specify conceptual work products (CWP), which in turn enables model checking for interactive HIT systems. The technique, presented in this paper, builds on earlier work that innovated CWP specifications as reference models to guide the design of interactive systems with workflows performed by human and machine agents 9. Early successes solved difficult design problems in aircraft schedules that integrate flying missions with aircraft maintenance 8, online self-support for users of personal computers with software problems 10, case management for MS outpatients 11, and NASA’s planning system to coordinate maneuvers of the International Space Station 5.

We introduce a new evaluation criterion for model checking interactive HIT systems using the CWP it is supposed to produce. A CWP can be specified using standard domain modeling techniques 7 for a clear description of a valid starting state, a valid ending state, and intermediate states that describe its acceptable evolutions. Importantly, the CWP abstraction tells us what must be accomplished by an HIT system independently of how it is accomplished. Rather than checking the state of a machine, the focus is on the state of the CWP as clinicians and their computing devices transform it to its goal state. Intuitively, this new approach shifts verification focus to the work intended to be accomplished by an interactive HIT system by proving that it accomplishes the work declared in the CWP.

To date, much of the research with model checking in health care has focused on formalizing workflows and proving temporal properties on workflows 2,3,19 or on developing new modeling languages to describe workflows and human machine teaming in those workflows 6, 12. A key principle of the new approach is that the CWP has states that are defined by combinations of attribute values that are independent of the means to attain them in the
workflow. This allows established model checking to be extended verify workflows that have both human-performed tasks and machine-preformed functions with vastly different capabilities in the same model. This new capability has potential to bring new levels of safety and reliability to HIT systems.

In the remainder of this paper we give an overview of the new method, illustrate the new techniques for modeling workflow and CWP with a small, interactive system, and describe a proof-of-technology study of the new model checking approach. We conclude by discussing the benefits to safety of HIT and to software engineering more generally, and outlining the next steps for research.

**Overview of the New Method**

We conducted a study of how to extend established model checking technology to exploit CWP as the evaluation criterion. Our objective was to determine whether CWP could serve as the evaluation criterion without making major changes to model checking technology. Our initial study assumed normative performance on all tasks by both human and machine performers. Our rationale was if a system could not check out under ideal conditions it did not merit further checking until it was improved. The functional components of the new approach in the study are shown in figure 1.

![Model Checker Components with CWP](image)

As shown in figure 1 the components have two main groups: Product & Process Models of the interactive system on the left, and the Verification Components on the right. The components for verification are widely recognized as well-established in effective practice for electronic and for software such as distributed systems. These verification components prove whether a transition system implements a specification.

The Process & Product Models on the left are two complementary innovations: for representing the products of the conceptual work of clinical care; and the workflow processes performed by clinicians and computers that create those products.

1. **MATHflow** is a graphical process diagramming tool that integrates models of clinical workflow with the use and change of information resources. There is growing interest in process models of workflow to analyze and design HIT. MATHflow is a partial implementation and an extension of the OMG’s recent standard for Business Process Modeling Notation (BPMN), which was created to support the development of software requirements for people working in groups that are supported by computing. As shown in figure 4 the MATHflow extension adds a properties editor for each task in a workflow to enter attributes to represent the information the task uses, the containing resource where each was accessed, and the destination information resource if the task changes any attribute values. The associations among tasks, attributes, and information resources are automatically recorded in MATHflow’s information dictionary.

2. In the terminology of knowledge representation a CWP is a declarative specification, as opposed to procedural specification of how-to do something. Rummelhart and Norman were among the first to report scientific
evidence that knowledge of what is learned and stored differently in human memory than the knowledge of how-to do things. CWP complement and complete procedural workflow models of care by making their purpose clearer and more meaningful. Generally, a workflow is constrained by the product it is supposed to produce because it implies so much about the activities and resources that are needed in the workflow. In the terminology of BPMN a CWP is a specification of a token, i.e., the entity that flows through a workflow process of task activities that operate on it to change its state. We use the term “conceptual work product” to mean the entity as it evolves through all its intermediate states to reach its goal state. In task-analytic methods such as TURF, a CWP corresponds to the work domain ontology 23. Complex CWP may require formal ontology modeling languages, such as OWL 21, while simpler CWP can be specified with standard declarative modeling languages such as a UML class diagram 7. In methods that use a domain model a CWP would be a key part of it. Figure 5 shows the class and state diagrams of the CWP for the proof-of-technology study.

The Subject System
The interactive system selected for model checking was Priority Contact for doctor-patient communication on test results and treatment plans for chronic disease. When patients have complex conditions doctors need to have real-time discussions with them about test results, diagnoses, and treatment options. Only about 10% of patients in this health system use internet, so phone is the modality of choice. Doctors also use the discussion to determine whether patients who are already dealing with complex treatments and medications will be able to follow the plan. An earlier version of Priority Contact is described in detail in Butler, et al.12.

Priority Contact illustrates the benefits of information technology for telemedicine that integrates desktop with phone technology, and the participation of clinician and patient. It also illustrates the liabilities of systems that involve remote, asynchronous participation that is supported by computing. The Priority Contact system depends on people and computing agents performing their tasks without the visibility that is more available when they are done manually. If any should fail there could be risk to patient health and safety that might go undetected.

The Workflows of Priority Contact
Figure 2 shows a screen shot of the MATHflow12 model of Priority Contact workflows that correspond to four levels of urgency, which were selected by the doctor in a previous task to initiate a contact plan. The workflows were designed to carry out their respective contact plans for:
1. The patient has a life-threatening condition - immediate contact must be made with the patient,
2. A life-changing diagnosis – a real time discussion should take place within a few days to begin major treatment,
3. A routine change to treatment – a telephone appointment should be offered to the patient to take place within fourteen days,
4. No change in treatment is required- the current plan should continue so a discussion appointment is optional to informing the patient of the test results.

MATHflow uses standard BPMN shapes to model the workflows of the four level of urgency with flows (arrows), decision gates (diamonds), tasks (plain rectangles) and sub-processes (rectangles with arrow head and cross at the bottom.

351
As shown in figure 3, sub-processes can, in turn, be opened to contain workflows of the same shapes. Figure 3 shows the contents of the sub-process Make Phone Clinic Appointment P3 that are displayed when it is opened. In addition, the workflow has swim lanes for graphically representing the task responsibilities of Patient and of Priority Contact as they interact to schedule a date and time for a telephone appointment for the patient and the doctor. The performer of a task is also recorded as one of its properties.
Figure 3: Modeling Human and Computer Interaction

Figure 4: Modeling the Use and Change of Information Resources
Figure 4 shows how the use and change of information is modeled as a property of the task PC writes timeDate to doctor’s schedule P3. The editor on the left is to record the information attributes that are used by the task, and the information object where it was accessed, e.g., patientName was accessed in the object Contact Plan. On the right a similar editor records any attribute whose value was changed and the destination object. In this example the task resulted in a value for confirmedAppointmentDateTime being entered to the Physician schedule object.

The MATHflow language treats information as a resource that is used or changed in tasks. In this manner any type of agent that performs a task, whether human or machine, can use information or change its value. This feature of MATHflow allows tasks to change the values of attributes that define the states of a CWP, regardless of whether they are performed by human or machine. This is a key principle for model checking on interactive systems whose workflows have both human and machine performers. The use and change of information in tasks is automatically recorded in MATHflow’s information dictionary. As a MATHflow model is incrementally developed the dictionary stores the associations among tasks, information attributes, values, and the objects where information is accessed.

The Conceptual Work Product of Priority Contact
As shown in figure 5 the CWP of Priority Contact is a Contact Plan represented with two diagrams. At the top the Contact Plan object is modeled as a UML class diagram. The bottom diagram is a UML state diagram.

Figure 5: Priority Contact’s Conceptual Work Product
Together, they are sufficient to specify the CWP and its states that the workflow must accomplish. The specification is independent of any given workflow, technology, or even cognitive strategy. Thus it can be used to verify other versions of the system that make the claim of satisfying the same requirements.

The Contact Plan is made up of several objects. At the center is a Conversation that is needed between Doctor and Patient. The Patient has Disease. The Doctor has Treatment Plan based on the Test Report. The Conversation is about the Treatment Plan. The model includes not only key attributes that define part of the actual state of the contact plan such as priority, launchTime, and resolveTime, but also doctor information, patient information, disease, treatment plan, etc. that is necessary information to accomplish the Contact Plan. All of this auxiliary information is read-only, and does not distinguish the actual state of the CWP at any given point in time and as such that information, with its various relationships to other information, can be abstracted in the verification model.

The states in the diagram at the bottom of figure 5 are defined by rules that use the key attributes. For example, the Contact Plan is in the Launched state when the attributes launchDate and launchTime have non-null values, but all other key attributes are still null. The Contact Plan can reach the Resolved state by either Conversation in progress for priorities 1-2, or Appointment scheduled for priorities 3-4. A Contact Plan may reach No longer needed if the patient somehow is no longer in MyCare of the Doctor.

**Model Checking**

The use of CWP for verifying interactive HIT systems begins by modeling the interactions of the clinician, patients, and Priority Contact as workflows in the MATHflow tool. The state diagram in the CWP is the actual specification for the interactive HIT. It defines the legal evolutions of the CWP from its initial state to its allowed final state. Any workflow must follow these legal evolutions when updating the state of the CWP. Model checking then follows every possible path in the workflow looking for bad sequences that do not reach one of the satisfactory sequences of the states that are required in the CWP specification. Sequences that violate the specification are considered unsafe because the planned progression of care does not happen, and it may go unnoticed. The rest of this section better defines model checking and describes in more detail the results of model checking the Priority Contact system.

Model checking works by constructing an exhaustive proof to show that a transition system implements a specification. Modeling any system requires the definition of all events in the system; these events are sometimes referred to as atomic propositions. A model produces a set of sequences of atomic propositions that represent the behavior of the system. Typically the model is described in some high-level language, and the semantics of that language yields sequences of atomic propositions that represent the language of the model as shown in the Venn diagram in Figure 6.

![Figure 6: Venn diagram illustrating model checking where the universe is sequences of atomic propositions in the model.](image)

The specification is also a language over atomic propositions that precisely describes sequences of interest. This language is usually presented in logic \cite{LTL, CTL}, and that logic describes sequences of atomic propositions that represent the correct language for the system. The inversion of this language creates a language of bad sequences that are not allowed by the specification. In the Venn diagram, this is the region labeled Universe minus Specification.

Model checking is the process of proving that the intersection of the language defined by the model and the language defined by the inverse of the specification is empty: no words are common to the two languages. If such is
not the case, then model checking provides a counter-example: a sequence produced by the model but not allowed by the specification.

The use of CWP for verifying interactive HIT systems is straightforward conceptually. The interactions of clinician, patient, and Priority Contact are modeled as a workflow in the MATHflow tool. The CWP is the specification. The atomic propositions for the language of the model and specification are the state labels in the CWP state diagram. For the Priority Contact model, these are: Launched, Conversation in Progress, Appointment Scheduled, Resolved, and No longer needed. The workflow model evolves the CWP from the initial Launched state and thus produces sequences of atomic propositions. These sequences are the language of the model. Any sequence from the initial state produced on any path in the workflow is part of that language.

The specification contains those sequences allowed by the state diagram in the CWP that start in the Launched state and end in the No longer needed state and only follow transitions defined in the state diagram. Sequences not in the language of the specification represent bad sentences that the model must never produce. For the priority contact example, bad sequences are those that do not end in the No longer needed state or those that make illegal transitions such as moving from the Launched state directly to the Resolved state.

Model checking follows every possible path in the workflow looking for bad sequences. If such a sentence exists, then the language generated by the model is not contained within the language defined by the specification, and the discovered bad sentence is reported as a violation. In general, model checking does not enumerate all bad sequences (if any exist) as only one is required to prove the relationship between the language of the model and the language of the specification.

The SPIN model checker is used in this work. The translation of the CWP and MATHflow model to the SPIN input language is not explained here but there is a fairly direct mapping from MATHflow tasks to processes in Promela. Most of the complexity in the translation is in sequencing and synchronizing the processes. The output from SPIN when no error is found is a coverage summary of the input model that indicates parts of the model that are not activated in the model checking. This inactivation is a result of not providing sufficient input to the model. For example, in the Priority Contact if a P4 no change lab result is never generated, then that portion of the model is not explored in model checking and is reported as uncovered. The output form SPIN when an error is found is a trace through the input file and a trail file that SPIN is able to replay in simulation mode for debugging purposes. If desired, then SPIN can be configured to return all errors and not just the first discovered error.

The rest of this section is devoted to the results of model checking the Priority Contact example. The SPIN model checker discovered errors in the workflows of all of the priorities. To bring this result into context, only one of the errors was seeded purposely by the modelers. The other workflows were believed to be correct.

Figure 2 is the top-level workflow of the different levels of priority in the Priority Contact system. Using this figure as a reference, a brief description of the discovered errors follows:

- **P1 Life Threatening**: in the PC carry out contact plan P1 process a gateway had no outgoing edge and was essentially disconnected from the rest of the model. As a result, any path leading to that gateway left the CWP in an invalid end state. The error was corrected by connecting the gateway to the remaining flows.
- **P2 Life Changing**: this error was seeded by the modelers. In the Doctor inform patient of results process action to move the CWP into the No longer needed state was removed by the modelers.
- **P3 routine change**: the modelers had included a Preferred mode of Communication gateway where the user is able to designate Mail, Auto-call, or MyHealthEvet. In the cases of Mail and MyHeathEvet, the tasks failed to move the CWP out of the Launched state. The error was corrected by removing the gateway as these options were hold-overs from a prior design that the designers had failed to remove.
- **P4 no change**: aside from the legacy gateway carried over from that prior design that was identical to the P3 routine change error, this workflow failed to move the CWP out of the Launched state. Neither the Make phone clinic appointment P4 nor the Discuss no change processes recorded the scheduled appointment, conversation, or resolution of the priority contact on any path. The error was corrected by modifying the tasks in Make phone clinic appointment P4 and Discuss no change. Tasks in these processes now change the state of the CWP indicating when the appointment is confirmed, the conversation takes place, and the doctor marks the contact as resolved.

The errors, though seemingly trivial, persisted through several reviews of the model, indicating the value of model checking if for nothing more than a sanity check on the connectivity of the model and the basic evolution of the CWP by the workflow.
Conclusion

The meaning of the model checking result is that the MATHflow model of the Priority Contact system implements the CWP under all possible inputs and under all possible ways for the actors to communicate. A faithful implementation of the MATHflow model would be correct relative to the intended work. The more general meaning is that a CWP can serve as a common entity which both clinicians and computers can change, thus allowing model checking technology to evaluate workflows that have both human-performed tasks and machine-preformed functions with vastly different capabilities in the same model. The MATHflow model and the CWP model could readily be translated into the input needed to apply established model checking tools. Extending model-checking to interactive health IT systems has great potential to advance HIT design and address safety in a powerful, formal method. To our knowledge no other approach can formally evaluate how well HIT is integrated with the fundamentally important manual activities of care. Before a system is actually built model checking can either prove that an interactive system can do the job it is supposed to do, or it identifies failures must be addressed.

Model checking proves a transition system implements a specification. To date, much of the research with model checking in health care has focused on formalizing workflows and proving temporal properties on workflows. The approach in this paper shifts focus to the work intended to be accomplished by a workflow. It uses model checking to prove that a workflow is capable to accomplish the work product that is specified in the CWP. Also, unlike prior work, this approach to model checking tries to use the CWP state diagram directly to prove the correctness of the system. This approach is in direct contrast to where the system specification is expressed in terms of formal logic.

Health IT is less mature than other established domains for system design. Absent a clear CWP specification and a technique for verifying at least partial algorithm correctness for producing a CWP, there is no basis to evaluate whether the workflow of a proposed HIT system is capable to accomplish the CWP. Further, it is not meaningful to compare systems on safety-related qualities such as usability if they have not been verified to produce equivalent work products. Consequently, HIT development may struggle with fundamental design decisions, such as how to achieve safe care improvements via appropriate allocation of functionality among clinical personnel and computing. Conversely, it is far more meaningful to compare workflows that have been verified as comparable for a variety of qualities such as reliability, usability, or efficiency, thus giving designers great latitude to find better solutions. Further, definitions of the checks could be instrumented as part of the HIT software to continue monitoring safety-related conditions after a verified system has been deployed. These capabilities should be important steps towards realizing the great potential of HIT.

Future Work

There are many directions opened up by the capability to verify socio-technical systems for healthcare. The new method first evaluates how the system will behave under idealized conditions, e.g., networks do not fail, and people show up for work on-time. We aim to investigate how the assumptions of the idealized system can then be relaxed to explore fault tolerance, and support designing to eliminate or mitigate the risks.

There are three critical pieces of future work related to the model checking step, itself: automation, scaling to large systems, results presentation, and workload analysis. The first step automates the process of extracting a Promela model from the MATHflow and CWP. It is critical for case studies on complex HIT systems. The second step characterizes the model checking on larger systems to understand its limits, and explores various techniques to mitigate state explosions including abstraction. The SPIN model checker typically scales to systems with millions of states, and it is hoped, but not yet known, that it is sufficient to verify interesting HIT processes. The third step relates directly to the usability of the model checking results. Any verification counter-example (i.e., a path that does not accomplish the work), must be mapped back into the MATHflow model and illustrated in MATHsim for the user to diagnose. We also plan to investigate how ontology modeling technology, with automated inferencing, can assist formally defining and scoping key CWP for clinical healthcare. These ontology models could provide a reusable inventory of key conceptual work products of clinical care to support health IT developers.

Acknowledgements

This research was supported by Grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects (SHARP) from the Office of the National Coordinator for Health Information Technology, and also by grant number R01HS021233 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.
References


Employing complex polyhierarchical ontologies and promoting interoperability of i2b2 data systems

James Richard Campbell MD, Walter Scott Campbell PhD, Hubert Hickman MS, Jay Pedersen MS, James McClay MD

1University of Nebraska Medical Center, Omaha, NE; 2Nebraska Medicine, Omaha, NE, 3University of Nebraska at Omaha, Omaha, NE

Abstract

I2b2 is in widespread use for managing research data warehouses. It employs reference ontologies as a record index and supports searching for aggregate cases using a pattern match operator on ASCII strings representing the node traversal from root to concept (PATHs). This creates complexities in dissemination and deployment for large polyhierarchical ontologies such as SNOMED CT. We hypothesized that an alternative approach employing transitive closure tables (TC) could lead to more accurate, efficient and interoperable search tools for i2b2. We evaluated search speed, accuracy and interoperability of queries employing each approach. We found both TC-based and PATH-based queries to produce accurate results. However, we observed that TC-based queries involving concepts included in large numbers of paths ran substantially faster than PATH-based queries for the same concept. Oracle query plan resource estimates differed by one to three orders of magnitude for these queries. We conclude that a simplification of dissemination tools for SNOMED CT and revision in the metadata build for i2b2 can effectively employ SNOMED CT with increased efficiency and comparable accuracy. Use of transitive closure tables in metadata can promote network query interoperability.

Introduction

Slowly and with little fanfare, reference terminologies and advanced ontologies supporting structured clinical data are reorganizing and enhancing electronic health records (EHR) and providing advanced data structures in support of clinical decision making. This has been accelerated in the US by coding requirements built into the Meaningful Use program by the Office of the National Coordinator for Health Information Technology. EHR vendors are struggling to understand and appropriately deploy these technologies and the research and public health communities have scant understanding of their importance or technical underpinnings.

With the advent of ambitious programs to develop and support large clinical research networks and the appearance of tools for big data bases, program administrators have little to guide them in the effective deployment of ontologies within data management systems. Arguably one of the greater challenges is SNOMED CT (SNCT) which is the most comprehensive healthcare ontology with a concept space of more than 302,000 elements. This large ontology can be difficult to effectively integrate into research data management tools. The September 2014 release of US edition of SNOMED CT is published as UMLS data sets of 3.05 Gbytes. The release format is an RF2 dataset specified by the IHTSDO but does not include any software or detailed guidance for best practice deployment into a system such as i2b2.

SNCT employs a concept model enforced by a description logic classifier which specifies the relationships and attributes to be used for concept definition in each SNCT domain. SNCT employs polyhierarchy (Definition: a hierarchically organized set in which a node may have more than one parent) in the ontology which means that the larger domains such as 404684003|Clinical finding (finding)| and 71388002|Procedure (procedure)| may have hierarchies of 30 generations in depth and many supertype concepts (Definition: X is the supertype of Y if a directed path of IS_A relationships can be traced from X to Y). Counting the possible paths from the root to a leaf concept, required for i2b2 metadata build, can be daunting since polyhierarchy expands the combinatorial possibilities. In the 2014 release of the US edition of SNCT there are up to 7802 paths (312411001|Carcinoma of the hepatic flexure (disorder)|) and an average of 11.1 paths per concept.

A research data management toolset in common use is i2b2. i2b2 software was first developed in 2004 when state of the art for terminology standards was ICD-9-CM and CPT-4. In contrast to SNCT, both are monohierarchies of modest size. (Definition: A hierarchically organized set in which any node may have at most one parent) i2b2 employs a star schema information model with four reference dimensions and a central fact table. i2b2 supports aggregation or subsumption queries (Definition: a subsumption query for concept X seeks to enumerate all instances from the fact table using the SQL pattern

359
match operator (LIKE) and an exhaustive set of PATHs from root concept to the node. For example, one possible path for the SNCT concept 128192002|Chronic disease of cardiovascular system (disorder)| is 404684003|64572001|276243003|128292002. Figure 1 shows the SNCT hierarchy for this concept and the sequence of nodes in path number one which illustrate the path computation. In total for this concept there are 4 possible paths.

Transforming the NLM release of the US edition of SNOMED CT into i2b2 metadata is difficult and time consuming, resulting in files of 4.984 Gbytes; these are troublesome to manage and share between collaborators seeking to standardize their research data. Distribution of i2b2 metadata files from the National Center for BioOntology using web services is available for SNOMED but reported to be seldom employed.

Importantly in this era of evolving national research networks, the i2b2 SQL query generated by the software is unique to the path selected by the user. It is only interoperable with other sites if their metadata deployment is exactly identical. To its credit, i2b2 manages such queries as efficiently as was considered possible in the age of ICD-9-CM.

Figure 1. SNOMED CT path computation

Extending previous work of Harris¹¹, we hypothesized that deployment of a revised i2b2 search algorithm for SNOMED CT employing simple transitive closure tables for subsumption queries would be faster and as accurate as conventional i2b2 methods. (Definition: A transitive closure set for any concept X in a IS_A hierarchy is the set of nodes which includes X and any node Y which has a path to X by following one or more directed ISA relationships¹² ) We further expected that a compact piece of software could install the SNOMED CT metadata using only NLM data sets. We believed that i2b2 queries employing transitive closure would be interoperable across i2b2 installations with metadata build which could be tailored for local needs.

Methods

We installed i2b2 version 1.7 in our research data warehouse on a Dell Poweredge R820 server with 16 cores and 128GB RAM; 2X744GB SAS solid state drives; 14 TB SAS 10K drives; supporting Oracle 11g.

We obtained the September 2014 RF2 release of SNOMED CT, US edition, from the UMLS Terminology Services site¹³ of the National Library of Medicine. We extracted the set of all active SNOMED CT concepts, relationships and descriptions from the snapshot version following the file documentation provided in the Technical Implementation Guide⁹. All concepts pertinent to this release were identified with moduleID values of 900000000000207008 and 900000000000012004 (International release) and 731000124108 (US-edition concept)¹⁴. We extracted the hierarchical (IS_A) relationships from the release files and developed a Python program that used those to prepare i2b2 metadata. First we identified the top concept for all SNCT hierarchies as first order subtypes of 38875005|SNOMED CT Concept|. We then calculated i2b2 metadata paths for SNCT employing an algorithm which is a modification of the standard depth-first search algorithm for graphs¹⁵. From the path search the program further prepared a transitive closure table for all of SNCT. A portion of the table is displayed in figure 2. The i2b2 metadata we prepared was employed in our i2b2 client as two versions: one using a conventional i2b2 PATH search and the Oracle ‘LIKE’ operator for computation of subsumption, and a second using a Transitive closure (TC) table and the ‘~’ operator. Figure 3 compares the metadata tables from i2b2 for the two search protocols. We created the
two versions of metadata to prepare the test environment for our research. i2b2 allows development for integration of ontologies. Historically, while most i2b2 metadata sets have been based upon concept paths, the flexibility of

![Fragment of transitive closure table](image)

i2b2 allows the development of other types of metadata such as dimension based metadata. Integration of the transitive-closure metadata table allows the i2b2 search engine to replace path-based calculation of subsumption with simple parent-child based subsumption using the TC table. Figure 3 shows the important columns from the two examples of i2b2 metadata. The first examples is transitive closure metadata, and the final column shows the traditional path based metadata. Note that we have created a concept_dimension_2 table analogous to concept_dimension that we use to isolate our testing with the traditional i2b2 style of metadata.

![Metadata attributes](image)

To compare accuracy and efficiency of the transitive closure methods, we developed a set of queries to be run using PATH-based logic and again using TC-based logic. The order in which the query logic was employed for each testing pair was randomized using a flip of the coin for each concept queried. The scope and complexity of each concept query varied from high-level concept such as 118234003|Findings by site(finding)| to mid-level concepts (128238001|Chronic heart disease (disorder)|) and low-level concepts (93030006|Congenital absence of spleen (disorder)|). We tested complex query performance by running queries that involved set calculations and disjoinoweeness, such as “count all patients with a chronic disease but not chronic cardiovascular disease”. We obtained database query resource requirements assessing query economy for each query using the “explain plan” SQL tool which provides an Oracle predicted relative resource value.

We further constructed an independent graph database and instantiated it with the fully classified SNCT model prepared from the same release files. We developed software to create the transitive closure table from the graph
database for validation of the i2b2 build. We compared the graph database transitive closure table to an instance developed with IHTSDO documented methods and validated the content of SNOMED CT. We copied all observation fact data from i2b2 to the graph database and reproduced the anonymized patient and SNOMED CT problem coding as an independent validation set. We then used the graph database to reproduce the concept queries that we employed for testing of TC and PATH query function and compared patient query sets for accuracy and reproducibility.

**Results**

The metadata implementation we developed employs the 20140901 RF2 snapshot release of SNOMED CT which is 503 MB of uncompressed terminology. The Python code produced a transitive closure table of 181 MB and a metadata set of 7.45 million rows at 4.984 GB. The metadata is complex due to SNOMED CT polyhierarchy and has from 1 to 7802 paths per concept. 16865 concepts of the 302,000 in the release have >=100 paths, most from 404684003|Clinical finding (finding)| hierarchy.

Table 1 summarizes the results of our controlled experiments with query execution. All queries retrieved identical patient sets regardless of query method, PATH-based or TC-based. We validated patient counts independently with queries of the graph database. We observed that subsumption query times for high-level concepts frequently ranged from 50% to 200% longer using the PATH-based approach. Query A in table 1 summarizes results for 404684003|Clinical finding (finding)| and query B 118234003|Finding by site (finding)| which are representative of these differences. Oracle relative resource estimates differed by up to three orders of magnitude in favor of the TC-based approach for these queries. We observed that query times involving concepts deeper in the SNOMED CT hierarchy were comparable between methods and differences between Oracle relative resource differences were negligible. Examples of these observations are reported in cases C, D and F. Queries incorporating disjunction for mid-level concepts also ran equally efficiently between methods (Table 1, query E). Surprisingly, we noted run time variability in PATH-based queries depending on browsing path chosen by the user. For example, see queries G and H in Table 1. In no case did PATH-based queries run faster that TC-based queries and summary statistics favored transitive closure procedures in an unpredictable manner that made non-biased quantitative assessment impractical.

Figure 4 recounts SQL generated by the i2b2 query engine to retrieve observation facts for query instances G and H from table 1. All three examples are searching for subsumed concepts of 13645005|Chronic obstructive lung disease (disorder)|. Queries 4a and 4b are both from path-based queries for this concept set but were generated when the user chose different navigation paths from the root to the concept during browsing for the concept in the SNOMED CT hierarchy. Query 4c is the SQL generated by the i2b2 query engine when employing metadata prepared for transitive closure searches as in figure 3. Note that SQL generated by the two PATH-based queries differs in the PATH segments chosen for searching.

![Figure 4. i2b2 SQL code fragments querying for 13645005|Chronic obstructive lung disease (disorder)| and 3b are path-based subsumption queries; 3c transitive closure-based subsumption query](image)
Discussion

We have documented our SNCT metadata build program for i2b2 and have published the code, transitive closure table and SNCT snapshot for PCORI sites which are using i2b2. Due to the relative simplicity of metadata creation from a transitive closure table for the use case of subsumption queries, we have discussed the publication of a transitive closure table for US Edition of SNOMED CT with NLM staff.

Deployment of the SNOMEDCT transitive closure tables allows construction of subsumption-complete subsets of SNOMEDCT. That is, with the subsumption logic embodied in the transitive closure table instead of explicit in enumeration of concept paths, subsets both in breadth and depth of the SNOMED CT ontology can be built. The build will find all subtype concept instances in the database, regardless of path build in the browser metadata set.

Requirements of the i2b2 metadata build are a real impediment to deployment of advanced ontologies such as SNOMED CT due to complexity and file size. Searches employing PATH-based metadata do not generate unique SQL statements in the setting of polyhierarchies and are a risk to the interoperability for queries to be shared within.

### Table 1. Query characteristics by method; run times and accuracy

<table>
<thead>
<tr>
<th>Query</th>
<th>Aggregation Logic</th>
<th>#paths containing target concept</th>
<th>Path-based query time in seconds (patient count)</th>
<th>Transitive Closure query time in seconds (patient count)</th>
<th>Graph database (Patient count)</th>
<th>Path queried</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>&lt;&lt;04684003</td>
<td>Clinical finding (finding)</td>
<td>1,489,885</td>
<td>43 seconds (442,271)</td>
<td>28 seconds (442,271)</td>
<td>442,271</td>
</tr>
<tr>
<td>B</td>
<td>&lt;&lt;118234003</td>
<td>Finding by site (finding)</td>
<td>3,029,152</td>
<td>37 seconds (392,771)</td>
<td>13 seconds (392,771)</td>
<td>392,771</td>
</tr>
<tr>
<td>C</td>
<td>&lt;&lt;128292002</td>
<td>Chronic cardiovascular disease (disorder)</td>
<td>492</td>
<td>1 second (5732)</td>
<td>1 second (5732)</td>
<td>5732</td>
</tr>
<tr>
<td>D</td>
<td>&lt;&lt;27624003</td>
<td>Chronic disease (disorder)</td>
<td>2318</td>
<td>2 seconds (51,546)</td>
<td>2 seconds (51,546)</td>
<td>51,546</td>
</tr>
<tr>
<td>E</td>
<td>&lt;&lt;27624003</td>
<td>Chronic disease (disorder) BUT NOT &lt;&lt;128292002</td>
<td>Chronic cardiovascular disease (disorder)</td>
<td>2318</td>
<td>2 seconds (45,814)</td>
<td>2 seconds (45,814)</td>
</tr>
<tr>
<td>F</td>
<td>&lt;&lt;930030006</td>
<td>Congenital absence of spleen (disorder)</td>
<td>28</td>
<td>&lt;1 second (2)</td>
<td>&lt;1 second (2)</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>&lt;&lt;13645005</td>
<td>Chronic obstructive lung disease (disorder)</td>
<td>1300</td>
<td>&lt;1 second (10,495)</td>
<td>&lt;1 second (10,495)</td>
<td>10,495</td>
</tr>
<tr>
<td>H</td>
<td>&lt;&lt;13645005</td>
<td>Chronic obstructive lung disease (disorder)</td>
<td>1300</td>
<td>1 second (10,495)</td>
<td>&lt;1 second (10,495)</td>
<td>10,495</td>
</tr>
</tbody>
</table>
a research network. In order to limit the size of metadata files, we had performed earlier experiments building only a limited number of paths for each SNCT concept. We verified that the strategy was untenable and produced erroneous responses to queries as well as dysfunction of the browsing environment. Therefore management of interoperability risk constrains i2b2 metadata since the metadata also supports the data browsing environment. By eliminating the need for a Concept_dimension table in the CRC cell and removing PATH data from the SQL query, a transitive closure metadata build is a step towards a flexible browsing interface, reduced i2b2 maintenance for terminology updates and improved interoperability of network query management.

We expect, but have not yet tested our results for consistency of i2b2 function on other SQL supported databases; namely SQL Server and Postgresql. We will be deploying our SNOMEDCT metadata and transitive closure table on each platform.

We conclude that the set of cases returned by the PATH and TC-based approaches were identical for each query. We confirmed the accuracy of these query results using an independent graph database representation of the data and different programmatic search engine. We included a complete and comprehensive enumeration of IS_A pathways in our build of i2b2 metadata because of the requirements of the i2b2 browsing software. Regardless of the specific path chosen by the user from the top-level concept, the query responses were identical for PATH and TC queries.

We consistently observed that speed and programmatic efficiency was as good or better with TC-based queries compared to conventional i2b2 PATH-based searches. As a puzzling observation, we found PATH-based queries of polyhierarchical concept data introduced unanticipated variation in query time response. While ASCII string length and the number of paths involving the target concept seemed to influence efficiency, we were unable to develop a quantitative model of query performance.

Conclusions

Deploying complex polyhierarchical ontologies in i2b2 can be a daunting task. We have documented and published one approach which has the virtues of employment of a US Meaningful Use2 standard ontology, modest file size and freely available software. Metadata maintenance workload is diminished since metadata tables are more compact. Extending the work of Harris11, we have confirmed that modifying i2b2 metadata build to employ transitive closure tables for support of subsumption queries leads to accurate results with retrieval efficiency as good or better than the current convention. SQL queries built from these modifications are devoid of browsing ontology references and are semantically parsimonious, promoting query interoperability.

References

Implementation and evaluation of a tele-education system for the diagnosis of ophthalmic disease by international trainees

J. Peter Campbell, MD, MPH\textsuperscript{1}, Ryan Swan, BS\textsuperscript{1,2}, Karyn Jonas, RN\textsuperscript{3}, Susan Ostmo, MS\textsuperscript{1}, Camila V. Ventura, MD\textsuperscript{4}, Maria A. Martinez-Castellanos, MD\textsuperscript{5}, Rachelle Go Ang Sam Anzures, MD\textsuperscript{4}, Michael F. Chiang, MD\textsuperscript{1,2}, R.V. Paul Chan, MD\textsuperscript{1}

Departments of \textsuperscript{1}Ophthalmology and \textsuperscript{2}Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR; \textsuperscript{3}Department of Ophthalmology, Weill Cornell Medical College, New York, NY; \textsuperscript{4}Altino Ventura Foundation, Recife (Brazil); \textsuperscript{5}Retina Service, Asociacion para Evitar la Ceguera en México, Hospital Luis Sánchez Bulnes, Mexico City, Mexico; \textsuperscript{6}Ospital ng Makati, Makati City, Philippines.

Abstract

Tele-education systems are increasingly being utilized in medical education worldwide. Due to limited human resources in healthcare in low and middle-income countries, developing online systems that are accessible to medical trainees in underserved areas potentially represents a highly efficient and effective method of improving the quality of the health care workforce. We developed, implemented, and evaluated an interactive web-based tele-education system (based on internationally accepted, image-based guidelines) for the diagnosis of retinopathy of prematurity among ophthalmologists-in-training in Brazil, Mexico, and the Philippines. We demonstrate that participation in this tele-education program improved diagnostic accuracy and reliability, and was preferred to standard pedagogical methods. This system may be employed not only in training, but also in international certification programs, and the process may be generalizable to other image-based specialties, such as dermatology and radiology.

Introduction

Medical education has undergone significant changes over the past decade with the integration of technology and online learning increasingly replacing physical lectures.\textsuperscript{1-5} In fact, so-called massive open online courses (MOOCs) are in some cases changing the entire paradigm of higher education away from on-site classroom learning.\textsuperscript{6-9} These online curricula have several advantages including standardization of content across large audiences, the ability to be utilized remotely, lower cost, and decreased dependence on local expert human resources for teaching.\textsuperscript{7,10} While in wealthier countries, this may change the form of knowledge delivery, in countries with less-developed healthcare and educational infrastructures, access to online tele-educational resources can dramatically change not only the form, but the quality and quantity of the educational content available to local trainees. Most importantly, tele-education systems allow for access to high quality education without an immediate need for an increase in local expert human resources, which may be non-existent or otherwise too busy to devote time to education.\textsuperscript{11,12} Therefore, the pedagogical argument for development of international tele-education systems is strongest for those specialties where the human resources are the most limited and/or the content is the most challenging to teach.

Retinopathy of prematurity (ROP) is a leading cause of childhood blindness worldwide.\textsuperscript{13} ROP care is clinically challenging and resource intensive. In the United States, despite a relative abundance of ophthalmologists and neonatologists compared to most of the world, there remain challenges to meeting the needs of ROP education as well as clinical care.\textsuperscript{14-16} Worldwide, this paucity of trained ROP providers is even more extreme in middle-income countries, where advancements in neonatal survival for preterm babies are resulting in increasing numbers of infants developing ROP.\textsuperscript{13} Moreover, accurate diagnosis of ROP can be challenging and many studies have demonstrated poor inter-examiner reliability, especially among providers with less experience, which raises serious concerns about the quality of care in medically underserved areas.\textsuperscript{14,16,17}

Tele-education can improve the exposure to and standardization of ROP education, and has potential to improve diagnostic competency of ophthalmology trainees worldwide. We developed an interactive tele-education system for ROP education and implemented this system at ophthalmology training programs in Brazil, the Philippines and...
Mexico. The purpose of this project was to determine the efficacy and trainee satisfaction of a remotely implemented ROP tele-education system. This system has potential to improve the quality and widespread delivery of ROP education and therefore indirectly (but significantly) improve the quality of medical care particularly in medically underserved areas throughout the world. Furthermore, this study demonstrates validation of principles that may be generalizable to other medical specialties, particularly ones that rely heavily on image-based diagnosis such as dermatology and radiology.

Study Domain

ROP is an ocular disease that develops in low birth weight and low gestational age infants, affecting approximately 14,000-16000 children in the United States per year.\textsuperscript{18} Diagnosis requires accurate recognition of disease, which has been classified by international guidelines established to triage these infants into one of several categories: no ROP, mild ROP, moderate ROP, and severe (treatment-requiring) ROP.\textsuperscript{19} Approximately 90\% of infants who develop ROP will have only mild or moderate disease and will require follow-up examinations every 1-2 weeks for the first few months of life.\textsuperscript{18,20} However, approximately 10\% of patients will progress to severe ROP. For these latter infants, treatment with laser photocoagulation or intraocular injection of pharmacological agents has been shown to successfully prevent blindness in most cases.\textsuperscript{18,20} Therefore, diagnostic accuracy of ROP that is high risk for requiring treatment is critical for preventing pediatric blindness.

Methods

The Weill Cornell Medical College (WCMC) and the Oregon Health & Science University (OHSU) Institutional Review Boards approved all aspects of the use and analysis of retinal images and educational material used in this study. Administration and analysis of the tele-education system was also reviewed by the WCMC Institutional Review Board, and was granted an exemption because it was considered research in an established or commonly accepted educational setting involving normal educational practices such as research on the effectiveness of instructional techniques, curricula, and instructional strategies. This study was conducted in adherence all federal and state laws and was in accordance with Health Insurance Portability and Accountability Act (HIPAA) guidelines.

Development of the Retinopathy of Prematurity Tele-education system

A secure web-based tele-education system was developed utilizing a repository of over 15,000 images captured from routine ROP exams. The system was coded in C# on the ASP.NET platform (Microsoft; Redmond, WA), and employed a SQL Server database. To facilitate software development and compatibility across different web browsers, the system was developed using an open source JavaScript wrapper (jQuery; Mozilla; Mountain View, CA). Modules were developed in both English and Portuguese.

Tele-education system content and procedures

Figure 1 demonstrates the procedure for the tele-education system, which consisted of four unique sections: (1) Pre-test examination; (2) ROP tutorial; (3) ROP training chapters; (4) Post-test examination. Clinical cases were presented in the Pre-test examination, ROP training chapters, and the Post-test examination. Each case included relevant clinical demographics and retinal images of both eyes. (Figure 1A). Trainees were required to enter a diagnosis for all image-based characteristics of ROP as defined by an internationally standardized classification system: plus disease (no, pre-plus, plus), zone (I, II, III), stage (1-5), category (no ROP, mild ROP, moderate ROP, severe treatment-requiring ROP), and aggressive posterior-ROP (APROP) for each individual eye. For the purposes of this study, the primary outcome measure was the overall disease category: presence of mild-or-worse, moderate-or-worse, or severe (treatment requiring) ROP.

Sections 1 and 4: Pre-test / Post-test. At baseline, and following the completion of the ROP training chapters each trainee completed a set of 20 clinical cases with various degrees of ROP. 16/20 cases (80\%) were unique and 4 (20\%) were repeated to measure intra-grader reliability. Completed cases could be reviewed but not changed during the testing and no feedback was provided.

Section 2: ROP Tutorial. Following the pre-test examination, trainees completed a tutorial on ROP diagnosis and management designed by the co-authors (RVPC, KEJ, SO, MFC). The ROP tutorial included didactic information on the different classifications of ROP and pertinent management considerations (treatment, follow-up time). When trainees were in between different chapters, the ROP tutorial was available as a reference for clarification, but the
tutorial was not available while trainees were in the pre-test or post-test examinations, or within a ROP training chapter.

Section 3: ROP Training Chapters. After completing the ROP tutorial, trainees were asked to complete five training chapters, each consisting of five clinical cases. Within each training chapter, trainees were given immediate feedback on their accuracy (Figure 1B) and targeted supplemental educational material was provided for any incorrect answers in four ways (Figure 1C): (1) Review of the image for which an incorrect response was provided; (2) Annotated images from the case that highlights the specific pathology; (3) Additional annotated images selected by the program from the database of validated images of other cases that highlight the specific pathology and; (4) Images that highlight the specific pathology corresponding to the (incorrect) chosen response. After reviewing the automated feedback based on their specific responses, trainees could then proceed to the next clinical case within the training chapter. After completing each chapter, trainees had the option to: (1) Proceed to the next chapter or section; (2) Review the ROP tutorial; (3) Review their responses and feedback from previous training chapters.

![Figure 1](image-url)

**Figure 1.** ROP Tele-education system procedures. (A) demonstrates progression through the course (pre-test examination, ROP tutorial, training chapters, and post-test examination), and displays a sample clinical case. In the training chapters, following the selected diagnosis, trainees were provided immediate feedback (B) and targeted supplemental tutorial and review (C).
Consensus Reference Standard Diagnosis

To ensure accuracy in the educational content, particularly because of potential variability in diagnosis among physicians, a reference standard ROP diagnosis was determined for each case by consensus of three experts and the clinical diagnosis. This was done using previously-described methods: (1) The clinical diagnosis (based on complete ophthalmic exam by an experienced examiner) was recorded; (2) Each set of retinal images was interpreted by 3 experienced readers using a web-based system; (3) The image-based diagnosis that was selected by the majority of image readers was then compared to the clinical diagnosis. When these two diagnoses are the same, it was defined as the consensus reference standard diagnosis. When the diagnoses are different, all of the data were reviewed by two of the investigators (RVPC, MFC) along with two study coordinators (KEJ, SO), and a consensus reference standard was determined. This consensus reference standard diagnosis was then used for the purposes of this current study.

Study Population

Ophthalmologists in training were recruited by the co-authors (KEJ, RVPC, RA, CVV, MAMC) from training programs in Brazil, Mexico, and the Philippines. Ophthalmology trainees in all years were allowed to participate in the ROP tele-education system. At one program, a control group was provided the pre-test examination, followed by the post-test examination to be taken at least one month after the pre-test examination (to prevent immediate recall), but was not given access to the ROP tutorial or ROP training chapters. At this program, trainees were randomized 1:1 to participate in the program as part of the control group or take the tele-education system as indicated in Figure 1.

Survey

A survey was designed using a publicly available service (http://www.SurveyMonkey.com). Six questions were adapted from existing psychometric instruments to measure trainees’ attitudes using a 5-point Likert-type scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree); one additional question allowed for free-text comments. Questions were reviewed by the authors for face validity and content validity and modified until all authors were satisfied with the survey instrument. Of the six questions, two assessed the trainees’ perception of their understanding of the diagnosis of ROP, three assessed the trainees’ attitudes toward preferred learning environment, and one assessed the trainees’ opinion of ease of use of the ROP tele-education system. All participants were invited to complete the survey.

Statistical Analysis

The main outcome measure was the post-ROP tele-education program sensitivity (compared to the pre-program sensitivity) of diagnosing any ROP (mild-or-worse ROP). The sensitivity and specificity of each disease category (mild-or-worse, moderate-or-worse, and severe) was measured before and after participation in the program. Sensitivities and specificities were compared using the paired t-test for each program and for all trainees as a single cohort, and by program, and statistical significance was considered to be a 2-sided P value <.05. For the training program with a control group, a secondary outcome measure included change in sensitivity of “active” education vs. control for each disease category. Data were analyzed using the R statistical software platform. Sensitivity and specificity values were calculated using the caret package. Based on the 4 cases that were repeated in both the pre-test and post-test examination, intra-grader reliability was evaluated using the kappa (k) statistic for chance-adjusted agreement in diagnosis. Cohen’s Kappa values were calculated using the irr package, and figures were generated using the ggplot2 package. A well-known scale was used for interpretation of results: 0 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.00, almost-perfect agreement. Descriptive statistics were used to summarize the trainees’ responses to the post-ROP tele-education program web-based survey. The non-parametric Wilcoxon signed-rank test was used to compare the trainees’ pre- and post-ROP tele-education program perception of their understanding of the diagnosis of ROP.

Results

Characteristics of Study Population

Table 1 summarizes key characteristics of the three training programs and study participants. 92 ophthalmologists-in-training were provided access to the ROP tele-education program among the three programs, and 8 completed the program: 27 at program 1 (with 26 randomized to control), 43 at program 2, and 11 at program 3. None of the training programs involved have a formal ROP curriculum to assess clinical competency of graduates, but all
programs had some exposure to ROP, and trainees were familiar with interpreting wide-angle retinal images such as those used in the program.

Table 1. Study population

<table>
<thead>
<tr>
<th>Program Characteristics</th>
<th>Program 1</th>
<th>Program 2</th>
<th>Program 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of trainees in program</td>
<td>58</td>
<td>54</td>
<td>11</td>
</tr>
<tr>
<td>ROP Curriculum (Y/N)</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Access to wide-angle camera (Y/N)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Premature infants per year examined by faculty</td>
<td>94</td>
<td>350</td>
<td>58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trainee Characteristics</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number completed program (%)</td>
<td>27*</td>
<td>43</td>
<td>11</td>
</tr>
<tr>
<td>Mean Age in Years (SD)</td>
<td>27 (1.2)</td>
<td>28 (2.5)</td>
<td>34 (2.8)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, n (%)</td>
<td>13 (48)</td>
<td>12 (28)</td>
<td>5 (45)</td>
</tr>
<tr>
<td>Female, n(%)</td>
<td>14 (52)</td>
<td>31 (72)</td>
<td>6 (55)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training year</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Year, n (%)</td>
<td>7 (26)</td>
<td>19 (44)</td>
<td>3 (27)</td>
</tr>
<tr>
<td>2nd Year, n (%)</td>
<td>8 (30)</td>
<td>15 (35)</td>
<td>2 (18)</td>
</tr>
<tr>
<td>3rd Year, n (%)</td>
<td>7 (26)</td>
<td>9 (21)</td>
<td>4 (36)</td>
</tr>
<tr>
<td>4th Year / Fellow, n (%)</td>
<td>5 (19)</td>
<td>N/A</td>
<td>2 (18)</td>
</tr>
</tbody>
</table>

*26 others participated as controls

Accuracy of ROP Diagnosis among Participants and Controls in the ROP Tele-education Program

Table 2 summarizes the pre- and post-ROP education program sensitivity and specificity for the diagnosis of ROP. The pre-test sensitivity of detecting mild-or-worse ROP was 81% (SE), which improved to 93% (SE) following the program, P < 0.001. There were statistically significant increases in sensitivity and specificity in all diagnostic categories. In subgroup analysis by program, these trends persisted with increases in sensitivity and specificity in each case.

Table 2. Accuracy of retinopathy of prematurity (ROP) diagnosis in the participants in the ROP tele-education program.

<table>
<thead>
<tr>
<th>ROP Disease Category</th>
<th>Sensitivity, % (SE)</th>
<th>Specificity, % (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Posttest</td>
</tr>
<tr>
<td>All programs (N=81)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild or worse</td>
<td>81 (2)</td>
<td>93 (1)</td>
</tr>
<tr>
<td>Moderate or worse</td>
<td>71 (3)</td>
<td>79 (2)</td>
</tr>
<tr>
<td>Severe (treatment requiring)</td>
<td>55 (4)</td>
<td>66 (4)</td>
</tr>
<tr>
<td>Program 1 (N=27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild or worse</td>
<td>83 (3)</td>
<td>95 (1)</td>
</tr>
<tr>
<td>Moderate or worse</td>
<td>67 (5)</td>
<td>81 (3)</td>
</tr>
<tr>
<td>Severe (treatment requiring)</td>
<td>54 (6)</td>
<td>71 (6)</td>
</tr>
<tr>
<td>Program 2 (N=43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild or worse</td>
<td>80 (3)</td>
<td>92 (1)</td>
</tr>
<tr>
<td>Moderate or worse</td>
<td>73 (4)</td>
<td>81 (2)</td>
</tr>
<tr>
<td>Severe (treatment requiring)</td>
<td>57 (6)</td>
<td>62 (5)</td>
</tr>
<tr>
<td>Program 3 (N=11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild or worse</td>
<td>78 (7)</td>
<td>90 (3)</td>
</tr>
<tr>
<td>Moderate or worse</td>
<td>73 (7)</td>
<td>82 (4)</td>
</tr>
<tr>
<td>Severe (treatment requiring)</td>
<td>48 (10)</td>
<td>66 (11)</td>
</tr>
</tbody>
</table>

ROP, retinopathy of prematurity; SE, standard error.
Table 3 demonstrates the change in sensitivity and specificity among the population of trainees randomized to participate in the program versus the control. In the control group there was no change in the sensitivity of detecting mild-or-worse disease (P = 0.708), the main outcome measure.

Table 3. Accuracy of retinopathy of prematurity (ROP) diagnosis in the trainees randomized to active tele-education and control at Program 1.

<table>
<thead>
<tr>
<th>ROP Disease Category</th>
<th>Sensitivity, % (SE)</th>
<th>Specificity, % (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Posttest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P value</td>
</tr>
<tr>
<td>Tele-education group (N=27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild or worse</td>
<td>83 (3)</td>
<td>95 (1)</td>
</tr>
<tr>
<td>Moderate or worse</td>
<td>67 (5)</td>
<td>81 (3)</td>
</tr>
<tr>
<td>Severe (treatment requiring)</td>
<td>54 (6)</td>
<td>71 (6)</td>
</tr>
<tr>
<td>Control (N=26)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild or worse</td>
<td>82 (4)</td>
<td>84 (3)</td>
</tr>
<tr>
<td>Moderate or worse</td>
<td>70 (5)</td>
<td>78 (4)</td>
</tr>
<tr>
<td>Severe (treatment requiring)</td>
<td>53 (7)</td>
<td>64 (7)</td>
</tr>
</tbody>
</table>

ROP, retinopathy of prematurity; SE, standard error.

Intra-Grader Agreement among Participants in the ROP Tele-education Program

Table 4 demonstrates the kappa statistics for intra-grader agreement among participants in the pre-test and post-test examinations of the ROP tele-education program. For the detection of mild-or-worse ROP among all participants (N=81), the kappa increased from moderate (0.57) to substantial agreement (0.8), with similar trends in every program and disease category. In the control group (N=26), the kappa also increased from 0.59 to 0.75, demonstrating improved intra-grader agreement with repeated testing over time, even without access to ROP tele-education program. 24 / 26 (92%) of control trainees waited the requested month between the pre-test and post-test examinations.

Table 4. Kappa statistics for intra-grader agreement participants and controls in the pretest and posttest of the retinopathy of prematurity tele-education program

<table>
<thead>
<tr>
<th>ROP Disease Category</th>
<th>All Programs (N=81)</th>
<th>Program 1 (N=27)</th>
<th>Program 2 (N=43)</th>
<th>Program 3 (N=11)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Posttest</td>
<td>Pretest</td>
<td>Posttest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild or worse</td>
<td>0.57</td>
<td>0.8</td>
<td>0.64</td>
<td>0.83</td>
</tr>
<tr>
<td>Moderate or worse</td>
<td>0.47</td>
<td>0.75</td>
<td>0.62</td>
<td>0.77</td>
</tr>
<tr>
<td>Severe (treatment requiring)</td>
<td>0.39</td>
<td>0.64</td>
<td>0.38</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ROP, retinopathy of prematurity.

POST-RETINOPATHY OF PREMATURITY TELE-EDUCATION PROGRAM SURVEY OF TRAINEES

31 of the 81 trainees (38%) who completed the ROP tele-education program also completed the post-program survey. Figure 2 shows the results of this survey. With a maximum score of 5 for each question, the average score for each question was calculated. The average score for question 1 (adequate understanding of ROP before the tele-education system) was 2.4/5, as only 5 (16%) of the respondents agreed or strongly agreed that they had an adequate understanding of ROP diagnosis before taking the ROP tele-education program. Following the program, the respondents’ score averaged 4.3/5 suggesting that they felt their knowledge significant improved through the program (signed-rank test; P<.01). 31 / 31 (100%) of the respondents agreed or strongly agreed that they had an adequate understanding of the diagnosis of ROP after completing the ROP tele-education program. The average score for question 2 (ROP tele-education was easy to use) was 4.2/5 as 29 (94%) of the respondents agreed or strongly agreed that the system was easy to use. 30 (97%) of the respondents agreed that the feedback at the end of cases was helpful (question 3). For question 6 (preference for interactive system over textbook learning), the average
score was 4.2/5 as 29 (94%) respondents agreed or strongly agreed that they learned more effectively with the automatic feedback opposed to textbook learning.

![Image of survey results](image_url)

**Figure 2.** Survey results of trainees who completed the retinopathy of prematurity (ROP) tele-education program (N = 31).

**Discussion**

**Summary of Key Findings**

We report the effective use of a tele-education system for teaching a clinically challenging ophthalmic disease in middle-income countries. We validate the concept that for image-based diagnosis, an internet-based tele-education system can be effective and preferred to more traditional educational models. The key findings from this study are:

1. This ROP tele-education system significantly improved the diagnostic accuracy and reliability of ophthalmology trainees.
2. Participants found the ROP tele-education system easy to use and they felt that they learned more effectively from ROP cases with automatic, targeted error-based feedback compared to a traditional textbook format.

**Implications for Retinopathy of Prematurity Care and Telemedicine**

These results demonstrate that a web-based tele-education system can be effective at improving the diagnostic accuracy for ROP for ophthalmology trainees in Brazil, Mexico, and the Philippines. Following completion of the course, trainees demonstrated 93% sensitivity at detecting mild-or-worse ROP, 71% sensitivity at detecting moderate disease, and 66% sensitivity at identifying severe (treatment-requiring) disease. To put this in perspective, these trainees outperformed a cohort of US pediatric ophthalmology fellows (85%, 53%, and 52% sensitivity for identifying mild, moderate, and severe disease, respectively) who completed ophthalmology training in the United States, but who did not complete this tele-education system. Additionally, participants in the tele-education program demonstrated improved intra-grader reliability of diagnosing ROP from “fair” or “moderate” to “substantial” agreement for all subtypes, including treatment-requiring ROP. These results are particularly significant in the diagnosis of ROP, in which despite the presence of international evidence-based, image-based guidelines, the literature consistently reports poor intra-rater and inter-rater reliability in ROP even among experts. As the routine care delivery model for ROP moves towards the use of telemedicine for screening and referral, the success of any telemedicine program will be dependent on the accuracy and reliability of the graders. Thus, though there remains room for improvement, particularly at recognizing moderate and severe ROP, as a pilot implementation of a tele-education system these results are extremely encouraging.

**Implications for International Tele-Medical Education**

The lack of human resources is perhaps the most critical factor limiting improvements in health care education and delivery in low and middle-income countries. In the US there is approximately one ophthalmologist for every 10,000 persons. In India, the number is closer to 1:100,000, and in most parts of Sub-Saharan Africa, it is less than 1:1,000,000. Compounding the human resource problem, among medical trainees in low and middle-income countries, access to quality medical education remains limited, and the ability to become well trained in the management of a complex disease such as ROP can be difficult. Even in the United States, ROP education has its challenges, as US trainees generally do not feel well qualified to perform ROP care independently without additional training. In middle-income countries where the number of cases of ROP is rising, the current medical education system has not been able to train enough qualified providers to manage emerging epidemic. Thus, a tele-education...
system that can be accessed from anywhere in the world that can demonstrate significant improvement in diagnostic accuracy and reliability has the potential to greatly impact the quality of ROP care, and mitigate the incipient rise in pediatric blindness in these countries. Moreover, these results suggest that a similar model could be used for other medical specialties that rely primarily on image-based diagnosis, such as dermatology, pathology, radiology, and cardiac electrophysiology and echography.

In addition to being utilized to improve diagnostic performance among trainees and practitioners, a similar web-based interactive system could be used to standardize and certify medical specialists for any of these specialties. The development of this web-based training system utilized a novel method for validating the reference diagnosis for each image against which trainees could be evaluated. Compared to traditional examinations with a limited number of testable images, and no intra-test interaction with the test-taker, there is tremendous potential for an interactive web-based system with thousands of validated images not only to assess clinical competency but to improve the accuracy and reliability of participants, and thereby improve the quality of “certified” care providers around the world.

Limitations
There are several limitations that should be mentioned. First, due to small numbers limiting statistical power, we were only able to randomize participants into a control group at one of the three training programs, and we were not able to include a control group based on more traditional pedagogical methods, such as textbook chapters and classroom lectures. Presumably, traditional methods of learning can provide similar results. That said, none of these programs currently have an established curriculum (nor do, for that matter, most US programs), and we have demonstrated that providing remote access to this tele-education system can markedly improve diagnostic performance, exceeding that of US graduates, with limited local effort or resources. Second, in this pilot study, we did not attempt to determine the common diagnostic errors that led to incorrect responses. Third, we are unable to estimate whether a longer program or some modification of this program would perform better, or a shorter program perform equivalently. Fourth, though we did include a Portuguese system for the Brazil group, we relied on the common familiarity with medical English in Mexico and the Philippines. It is possible that these results may not easily translate when multiple translations are required for other countries and languages. We plan to investigate these questions in future studies.

Conclusions
In summary, we demonstrate that participation in this ROP tele-education system can increase diagnostic accuracy and reliability among trainees in Brazil, Mexico, and the Philippines. Program participants found the web-based, interactive system easy to use, and preferable to traditional learning methods. These results provide evidence that tele-education may be an effective modality to improve diagnostic performance in ophthalmologic disease, and particularly in middle-income countries, where there is an emerging epidemic, tele-education may represent a key strategic intervention to improve the quantity and quality of ROP care providers without added burden to existing local human and educational resources. As telemedicine and tele-education approaches to clinical care and education become more mainstream, we believe that these results may be generalizable to other medical specialties which rely on image-based detection of disease using international standards.

References
6. Hossain MS, Shofiqul Islam M, Glinsky JV, Lowe R, Lowe T, Harvey LA. A massive open online course (MOOC) can be used to teach physiotherapy students about spinal cord injuries: a randomised trial. J Physiother.


Matthias Gamer, Jim Lemon and Ian Fellows Puspender Singh <uspender.pusp22@gmail.com> (2012). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84. http://CRAN.R-project.org/package=irr


Chiang MF, Wang L, Busuioic M, Du YE, Chan P, Kane SA, et al. Telemedical retinopathy of prematurity...

Acknowledgments

Supported by The St. Giles Foundation (RVPC); Grant R01 EY019474 from the National Institutes of Health (KEJ, MFC, RVPC, SO); The Bernadotte Foundation for Children's Eyecare (KEJ, RVPC); The iNsight Foundation (KEJ, RVPC); Unrestricted departmental funding from Research to Prevent Blindness, Inc, New York (KEJ, MFC, RVPC, SO); Novartis Excellence in Ophthalmic Vision Award – XOVA (KEJ, RVPC).
Representation of Drug Use in Biomedical Standards, Clinical Text, and Research Measures

Elizabeth W. Carter, MS1, Indra Neil Sarkar, PhD, MLIS1,3, Genevieve B. Melton, MD, PhD4,5, Elizabeth S. Chen, PhD1,2
1Center for Clinical & Translational Science, 2Department of Medicine, 3Department of Microbiology & Molecular Genetics, University of Vermont, Burlington, VT; 4Institute for Health Informatics, 5Department of Surgery, University of Minnesota, Minneapolis, MN

Abstract

Drug misuse is a prominent cause of morbidity and mortality in the United States. Recent focus on behavioral and social domains in the electronic health record (EHR) has highlighted the need for comprehensive examination of social history information, such as drug use. In this study, representation of drug use was examined in three types of sources: (1) standards from HL7 and openEHR, (2) clinical text from publicly accessible clinical notes and a local EHR, and (3) research measures from the PhenX Toolkit and CDE Browse. In total, 27 elements were identified across the examined sources, revealing a diverse set of values that were found to be associated with drug use type, frequency, method, time frame, and amount. The findings of this study provide insight into the representation of drug use information that may contribute to efforts for standardizing collection and use of these data to support clinical care and research.

Introduction

Morbidity and mortality from pharmaceutical, over-the-counter, and illicit drug misuse in the United States (U.S.) has increased in recent years, while drug poisoning has become a leading cause of injury deaths1–2. Opioid analgesic poisoning deaths tripled between 2000 and 20103. Opioids and benzodiazepines were the predominant pharmaceutical drugs leading to over 43,000 overdose deaths in 20134. In 2013, 8.8% of adolescents and 9.4% of adults in the U.S. were current (within the past month) users of illicit drugs5. Further, drug misuse has been well documented among youths6 and adults7–8 to be associated with comorbid medical problems and mental health disorders. The consequences of such comorbidities for youths can be substantial – including higher rates of treatment, social and academic problems, and suicide attempts6,7. In the general population, co-occurrence of mental health disorders with drug misuse has been documented for anxiety disorders (with marijuana)9,10 and suicide deaths (with misuse of prescription drugs) in U.S. veterans11. These findings reflect part of the impact of drug use in the U.S. and illuminate the need for standardized representation of drug use in behavioral data that may be used to inform patient care, clinical research, and public health policy.

In recent years, focus on standardized collection of patient information using the electronic health record (EHR) has intensified. The 2009 Health Information Technology for Economic and Clinical Health Act (HITECH) outlined goals for the adoption of EHRs for health providers, in parallel with the objective of “meaningful use” to improve patient care12. The Patient Affordable Care Act of 2010 further emphasized the detection of substance use problems and the integration of care with primary care providers13. The 2013 set of Meaningful Use Objectives from the Centers for Medicare and Medicaid Services contains one core objective for tobacco use (smoking status); however, there were no detailed objectives for other types of drug use14. A collaboration of the National Institute on Drug Abuse Center for Clinical Trials Network (NIDA-CCTN), National Cancer Institute (NCI), and Substance Abuse and Mental Health Services Administration (SAMSHA) supports the development of a standardized clinical quality measure (CQM) that may be considered for inclusion as a Meaningful Use core objective15. In 2014, the Institute of Medicine proposed a new set of standard measures for social and behavioral domains that included evaluation of the ‘Abuse of Other Substances’ domain16,17. However, due to the sensitivity and complexity of data collection for drug misuse, this domain was not included for further consideration for Meaningful Use Stage 3. Existing standards for collection of social history information (e.g., the HL7 implementation guides for Clinical Document Architecture (CDA) Release 218 and the ‘Substance Use Summary’ and ‘Substance Use’ archetypes in openEHR19) that may provide a possible way to represent drug use in EHRs have been assessed in prior work for capturing social history information in clinical notes20.

The EHR has the potential to serve as a powerful tool to support drug use screening, diagnosis, intervention, and treatment for primary care providers21. Wu, et al. examined the prevalence of substance use disorders (SUDs) and

376
patterns of comorbidities in adults by extracting patient information from the ‘Habit/Substance Use’ domain in the EHR at Duke University Medical Center. Prevalence of SUDs differed by sex, race and ethnicity, whereas comorbidities differed by race and sex, and were more prevalent among patients with SUDs than those without. This study provides supporting evidence for the use of EHRs for research to inform health care. Yet, further standardization of patient data is required to support information sharing among providers. Motivated by the 2006 National Institutes of Health (NIH) National Electronics Clinical Trials and Research projects, NIDA-CCTN led an effort to develop sets of consensus-based common data elements (CDEs) for substance use to be used within the EHR and other data sources. The NIH Common Data Element (CDE) Resource Portal promotes the use of CDEs through numerous initiatives, tools, and resources. Among these are the Cancer Data Standards Registry and Repository (caDSR) that incorporates cancer-specific CDEs from numerous sources; Grid-enabled Measures (GEM) that facilitates the development and sharing of common measures; and, Consensus Measures for Phenotypes and Exposures (PhenX) that provides consensus-based standard measures to be incorporated in genome-wide association and epidemiologic research studies. These sources contain CDEs (a single question denoting a fixed representation of a variable) or measures (a standardized instrument containing a protocol with one or more CDEs), which will henceforth be referred to as “research measures.”

With the increased attention to the importance of monitoring and understanding drug use and its impact, there is a need for research focused on comprehensive collection and subsequent use of this information. To address this need, the goal of this study was to examine the representation of drug use in multiple sources for informing EHR and standards development for guiding evidence-based patient care and ultimately improving patient outcomes.

Methods

Study design

Three types of sources were analyzed to identify elements associated with drug use and their corresponding values: (1) standards — HL7 CDA-based models and openEHR archetypes; (2) clinical text — drug use sentences in clinical notes from the publicly accessible MTSamples.com (MTS) resource and drug use comments from the social history module of the Epic EHR at the University of Vermont Medical Center (UVMMC); and (3) research measures — measures including one or more CDEs from the PhenX Toolkit and CDEs from caDSR using the CDE Browser.

Analysis of standards

The first phase involved exploring current standards for collection of social history information related to drug use. The HL7 Implementation Guide for CDA Release 2: IHE Health Story Consolidation, DSTU Release 1.1 was examined for elements within the Social History Observation template of the Social History section. Archetypes for ‘ Substance Use Summary’ and ‘Substance Use’ were also identified using the openEHR Clinical Knowledge Manager. Elements described within these standards were merged with a set of elements identified in previous work involving the analysis of social history information in clinical notes, social and behavioral information in public health surveys, and free-text comments associated with tobacco and alcohol use in the EHR. This combined set of 25 elements was used to create an initial set of annotation guidelines for analyzing the clinical text and research measures.

Analysis of clinical text

The second phase consisted of an analysis of clinical texts from MTS and UVMMC. For MTS, sentences describing drug use were identified within 491 clinical notes categorized as ‘Consult – History and Physical’, which have been used in prior studies, using the General Architecture for Text Engineering (GATE). This provided a set of 130 drug use sentences from 124 (25.3%) notes. At UVMMC, the Epic EHR social history module collects information on drug use using a set of structured fields for Status (‘Not Asked,’ ‘Yes,’ and ‘No’), Use/week, and Type (e.g., benzodiazepines, ‘cocaine,’ ‘marijuana,’ ‘heroin,’ and ‘opiods’) as well as a free-text comment field, which was the focus of this study. A random set of 50 drug use comments from January 2014 was first analyzed to determine if and how to enhance the annotation guidelines. For example, the original element Duration was found to be too broad and led to the creation of more specific elements Duration of Use and Duration Since Time Point. Two annotators (EWC and ESC) annotated a subset of 10% of the 130 MTS sentences and 450 UVMMC comments from March 2014 (representing the most recent comments for a random set of 450 patients), achieving an inter-rater reliability using Cohen’s kappa of 0.92 and 0.94 respectively.
The brat rapid annotation tool (BRAT)\textsuperscript{41,42} was then used to annotate drug use information in the 130 sentences and 450 comments (Figure 1) using the revised annotation guidelines. These guidelines define the elements (e.g., Status, Type, or Time Frame) as well as relationships between elements such as Type and Time Frame (e.g., ‘40 years ago’ => ‘heroin’). Further analysis involved extracting the annotations (representing element-value pairs) and generating statistics for each element and their corresponding values for each source of clinical text as well as the combined set of annotations for MTS and UVMMC. Using the combined annotations, those with similar meaning were grouped by a common word, phrase, or pattern. For example, the grouping ‘Cocaine’ represented the annotations ‘cocaine,’ ‘cocaine drug,’ ‘crack,’ and ‘crack cocaine,’ whereas patterned groupings such as ‘[#] [time unit] ago’ represented ‘days ago’ and ‘over a year ago.’

**Figure 1.** Annotation of clinical text using the brat rapid annotation tool.

**Analysis of research measures**

The third phase of the study involved analysis of the PhenX Toolkit (V5.7; accessed July 31, 2014) containing 21 domains with over 330 measures and the caDSR\textsuperscript{43} using the CDE Browser (V4.0.4; accessed July 31, 2014) providing access to 35 resources containing information about thousands of CDEs. A search for measures using the PhenX Toolkit was done by browsing the domains with a focus on the ‘Alcohol, Tobacco, and Other Substances’ domain, in addition to a search of all measures by using the following search terms: ‘substance,’ ‘illicit,’ ‘illegal,’ and ‘drug.’ The same set of search terms was used to identify measures in the CDE Browser. Cumulatively, an initial set of 111 measures was identified and further restricted by excluding measures that explicitly targeted substance abuse and cessation; emotional, physical or social repercussions of drug use or abuse; or, opinions and perspectives about use or abuse. This resulted in a final set of 40 drug use-specific measures containing 40 questions from CDE Browser and 10 drug use-specific measures containing 191 questions from PhenX Toolkit. For some measures, instructions and supplemental information were also included in the analysis. For example, the measure ‘Patterns of substance use – adults’ in the PhenX Toolkit included a supplemental drug card detailing over 150 drug names and each was designated as a value within the corresponding measure.

**Figure 2.** Analysis of questions and responses within research measures.

Annotation of the measures (Figure 2) was performed using the same guidelines for annotating the clinical text by designating words and phrases within questions as well as the responses to questions as values (each equivalent to one annotation), and then assigning a corresponding element to each value. A new element for Time Point (reference to a particular period of time) was observed within the questions of the PhenX measures and subsequently added to the guidelines. Prior to annotation of the two sets of identified measures, two annotators (EWC and ESC) evaluated
a subset of 10% of the measures from CDE Browser and PhenX Toolkit and achieved an inter-rater reliability using Cohen’s kappa of 0.94 and 0.95 respectively. Similar to the clinical text, further analysis involved extracting the elements and corresponding values and generating statistics for each source as well as the combined set of measures. Using the combined annotations, those with similar meaning were grouped by a common word, phrase, or pattern.

**Results**

The HL7 and openEHR standards, 50 measures from PhenX Toolkit and CDE Browser containing a total of 131 questions, and 130 MTS sentences and 450 UVMMC comments represented a set of 27 drug use elements. Table 1 summarizes the distribution of these elements across the standards, clinical text, and research measures. In addition, this table shows the number and proportion of annotations associated with each element for MTS sentences (465 annotations), UVMMC comments (1,205 annotations), PhenX Toolkit (2,019 annotations), and CDE Browser (343 annotations). Of the 25 elements from prior studies\(^{20, 36, 37}\) (indicated by ‘\*’ in Table 1), 24 retained their original meaning. *Duration* was further demarcated as *Duration of Use* and *Duration Since Time Point*, and a new element *Time Point* was added, which resulted in a total of 27 elements.

**Table 1**: Distribution of drug use elements across sources.

<table>
<thead>
<tr>
<th>Element</th>
<th>Standards (n=465)</th>
<th>Clinical Text (n=1,205)</th>
<th>Research Measures (n=2,019)</th>
<th>CDE Browser (n=343)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certainty*</td>
<td>35 (7.5%)</td>
<td>103 (8.5%)</td>
<td>1 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>Negation*</td>
<td>5 (1.1%)</td>
<td>1 (0.1%)</td>
<td>140 (69.0%)</td>
<td>20 (8.5%)</td>
</tr>
<tr>
<td>Temporal*</td>
<td>110 (23.7%)</td>
<td>54 (4.5%)</td>
<td>61 (3.0%)</td>
<td>17 (7.3%)</td>
</tr>
<tr>
<td>Start Date*</td>
<td>X</td>
<td>-</td>
<td>2 (0.2%)</td>
<td></td>
</tr>
<tr>
<td>Start Age*</td>
<td>X</td>
<td>-</td>
<td>3 (0.2%)</td>
<td></td>
</tr>
<tr>
<td>Quit Date*</td>
<td>X</td>
<td>-</td>
<td>5 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>Quit Age*</td>
<td>X</td>
<td>-</td>
<td>9 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>Duration of Use</td>
<td>-</td>
<td>-</td>
<td>9 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>Duration Since Quit*</td>
<td>-</td>
<td>-</td>
<td>11 (0.9%)</td>
<td></td>
</tr>
<tr>
<td>Duration Since Time Point*</td>
<td>-</td>
<td>-</td>
<td>3 (0.2%)</td>
<td></td>
</tr>
<tr>
<td>Time Frame*</td>
<td>6 (1.3%)</td>
<td>84 (7.0%)</td>
<td>174 (8.6%)</td>
<td>7 (3.0%)</td>
</tr>
<tr>
<td>Time Point</td>
<td>7 (1.5%)</td>
<td>48 (4.0%)</td>
<td>7 (0.3%)</td>
<td>3 (1.3%)</td>
</tr>
<tr>
<td>Method*</td>
<td>X</td>
<td>X</td>
<td>99 (21.3%)</td>
<td>147 (12.2%)</td>
</tr>
<tr>
<td>Type*</td>
<td>X</td>
<td>X</td>
<td>145 (31.2%)</td>
<td>327 (27.1%)</td>
</tr>
<tr>
<td>Subtype*</td>
<td>X</td>
<td>X</td>
<td>673 (33.3%)</td>
<td>152 (65.0%)</td>
</tr>
<tr>
<td>Amount*</td>
<td>X</td>
<td>X</td>
<td>16 (3.4%)</td>
<td>69 (5.7%)</td>
</tr>
<tr>
<td>Frequency*</td>
<td>X</td>
<td>X</td>
<td>5 (1.1%)</td>
<td>218 (18.1%)</td>
</tr>
<tr>
<td>Context*</td>
<td>X</td>
<td>X</td>
<td>5 (1.1%)</td>
<td>19 (1.6%)</td>
</tr>
<tr>
<td>Situation*</td>
<td>-</td>
<td>-</td>
<td>2 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>Location*</td>
<td>-</td>
<td>-</td>
<td>2 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>Subject*</td>
<td>-</td>
<td>-</td>
<td>37 (1.8%)</td>
<td></td>
</tr>
<tr>
<td>Change*</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Triggers*</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Evidence of Dependence*</td>
<td>X</td>
<td>X</td>
<td>21 (4.5%)</td>
<td>21 (1.7%)</td>
</tr>
<tr>
<td>Cessation*</td>
<td>X</td>
<td>1 (0.2%)</td>
<td>6 (0.5%)</td>
<td>9 (0.4%)</td>
</tr>
<tr>
<td>Other*</td>
<td>X</td>
<td>5 (1.1%)</td>
<td>20 (1.7%)</td>
<td>3 (0.1%)</td>
</tr>
<tr>
<td><strong># of Elements</strong></td>
<td>7</td>
<td>15</td>
<td>16</td>
<td>23</td>
</tr>
</tbody>
</table>

\* number of annotations; \* element derived from previous work\(^{20, 36, 37}\).

As reflected in Table 1, seven elements were represented in the HL7 implementation guide while the openEHR archetypes cumulatively included 15 elements. MTS clinical text and measures from PhenX Toolkit and CDE Browser contained information for 16, 18, and 15, elements, respectively. Clinical text from UVMMC contained the most diverse drug use information representing 23 elements. *Quit Age*, *Change*, and *Triggers* were only found in the openEHR archetypes whereas *Time Frame* (e.g., ‘last 6 months’), *Time Point* (e.g., ‘tonight’), and *Context* (e.g., ‘at
night for sleep’) were specific to the research measures and clinical text. *Situation* (e.g., ‘when going out to eat’) and *Subtype* (e.g., ‘pills: vicodin’) occurred only in clinical text.

Within the standards, values were provided for some elements. For example, in the openEHR archetypes, values for *Status* included ‘current user,’ ‘former regular user,’ ‘former occasional user,’ and ‘never user’ while *Frequency* values included ‘daily use,’ ‘weekly use,’ ‘irregular use,’ and ‘no use.’ In addition, data types were provided that may be used to infer values for some elements (e.g., Date/Time for *Start Date* and *Quit Date*). Analysis and grouping of values focused on six elements that were found to be the most frequent across the clinical text and research measures: *Type, Frequency, Method, Negation, Time Frame,* and *Amount.* Negation was the third most frequently occurring element and included six groups of values: ‘no,’ ‘none,’ and ‘without’ (78 total and 3 unique values in measures) and ‘denies,’ ‘does not,’ and ‘never’ (164 total and 10 unique values in clinical text). Tables 2-6 include the value groupings for the remaining top five elements and shows for each element: (1) total number of values, number of unique values, and number of groups; (2) total number of values per group, frequency of group among the total values for element, and number of unique values within the group; and (3) example values for the highest frequency groups per element.

### Table 2: Distribution of values for Type element. (total # of values per group; (frequency); [# unique values])

<table>
<thead>
<tr>
<th>Clinical Text</th>
<th>Research Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>472 Total Values; 109 Unique Values; 34 Groups</td>
<td>824 Total Values; 408 Unique Values; 203 Groups</td>
</tr>
<tr>
<td>Marijuana: 83 (17.6%) [11]</td>
<td>Drugs: 69 (8.4%) [16]</td>
</tr>
<tr>
<td>Cannabis</td>
<td>Any other drugs, drug products</td>
</tr>
<tr>
<td>Marijuana, marijuana, marijuanna, MJ</td>
<td>Drug non-medical use only</td>
</tr>
<tr>
<td>Medical marijuana, prescribed marijuana</td>
<td>Drugs not prescribed by a doctor</td>
</tr>
<tr>
<td>Caffein, caffeine</td>
<td>Cannabinoids, cannabis</td>
</tr>
<tr>
<td>Coffee, coffees, ice coffee, green tea, tea</td>
<td>Hash, hash oil, hashish</td>
</tr>
<tr>
<td>Cola, pepsi, soda, energy drinks</td>
<td>Marijuana, mary jane</td>
</tr>
<tr>
<td>Illegal drugs: 62 (13.1%) [10]</td>
<td>Cocaine: 47 (5.7%) [10]</td>
</tr>
<tr>
<td>Illegal drug, illicit substance</td>
<td>Angel dust</td>
</tr>
<tr>
<td>All illicit, illicit drug, illicits</td>
<td>Cocaine, coke, crack</td>
</tr>
<tr>
<td>Multiple illicit drugs</td>
<td>Cocaine in chunk form</td>
</tr>
</tbody>
</table>

Table 2 shows the distribution of values for the element *Type* among clinical text and research measures. Values in the group ‘Marijuana’ were among the most commonly observed at a frequency of 6.1% of 842 values (research measures) and 17.6% of 472 values (clinical text) with the values ‘marijuana’ and ‘cannabis’ observed in both sources. The group ‘Marijuana’ also illustrates the occurrence of misspellings and abbreviations among the clinical text as observed with ‘marijana,’ ‘marijuanana,’ and ‘MJ.’ The group ‘Caffeine’ was observed only in clinical text and included values representing coffees, teas, and soft drinks.

### Table 3: Distribution of values for Frequency element. (# values; (Frequency); [# unique values])

<table>
<thead>
<tr>
<th>Clinical Text</th>
<th>Research Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>223 Total Values; 81 Unique Values; 7 Groups</td>
<td>672 Total Values; 52 Unique Values; 7 Groups</td>
</tr>
<tr>
<td>Occasionally: 114 (51.1%) [13]</td>
<td>[#]: 576 (85.7%) [18]</td>
</tr>
<tr>
<td>Occasionally, off and on, on occasion</td>
<td>Exact number: e.g., 0, 1, 2</td>
</tr>
<tr>
<td>Every few months</td>
<td>Range: e.g., 1-2, 6-9, 20-39</td>
</tr>
<tr>
<td>Every once in a while</td>
<td>&gt;=&lt;: e.g., 100 times or more</td>
</tr>
<tr>
<td>Intermittent</td>
<td>Calculation: e.g., average # times</td>
</tr>
<tr>
<td>[#] [time unit] per [time unit]: 102 (45.7%) [54]</td>
<td>[#] [time unit] per [time unit]: 51 (7.6%) [22]</td>
</tr>
<tr>
<td>/day, daily</td>
<td>Daily, weekly, monthly</td>
</tr>
<tr>
<td>/week, weekly</td>
<td>1 or 2 days a week, 2x weekly</td>
</tr>
<tr>
<td>1-2 times per month, 3x a day</td>
<td>About once a day, / day</td>
</tr>
<tr>
<td>Every other day, per day everyday</td>
<td>&lt;= once a month, monthly or less often</td>
</tr>
<tr>
<td>Other: 5 (2.1%) [4]</td>
<td>Other: 12 (1.6%) [5]</td>
</tr>
<tr>
<td>Chronic</td>
<td>More often than prescribed</td>
</tr>
<tr>
<td>Minimal, not usually</td>
<td>Most frequently, usually</td>
</tr>
<tr>
<td>Part-time</td>
<td>Never</td>
</tr>
</tbody>
</table>
Values for the *Frequency* element were categorized into seven groups for the clinical text and research measures with the three most frequent groups represented in Table 3. The group ‘[#] [time unit] per [time unit]’ contained similar values for both sources including ‘daily’ and ‘weekly.’ In contrast, the most common *Frequency* group for clinical text was represented by the group ‘Occasionally’ while the measures reflected more precise values such as exact numbers (e.g., ‘0,’ ‘1,’ and ‘2’) or ranges (e.g., ‘1-2’ and ‘20-39’) in the group ‘[#].’

Table 4 shows that the top four groups for the *Method* element for clinical text and research measures are nearly equal with the exception that the value ‘intake’ appears in the clinical text whereas the value ‘taken’ is found in research measures. Values in the group ‘Use’ were represented in over 50.0% of the total values in clinical text and research measures. Within the measures, the group ‘Inject’ contained more descriptive values in the form of phrases such as ‘injected using needle’ and ‘injection into the skin,’ while clinical text contained more abbreviated values such as ‘inject,’ ‘boot,’ and ‘IVDU.’

<table>
<thead>
<tr>
<th>Table 4: Distribution of values for <em>Method</em> element. (# of values; (Frequency); [# unique values])</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical Text</strong></td>
</tr>
<tr>
<td>246 Total Values; 33 Unique Values; 12 Groups</td>
</tr>
<tr>
<td><strong>Use</strong>: 158 (64.2%) [7]</td>
</tr>
<tr>
<td>Us, use, used, uses, usage, using</td>
</tr>
<tr>
<td><strong>Intake</strong>: 32 (13.0%) [1]</td>
</tr>
<tr>
<td>Intake</td>
</tr>
<tr>
<td><strong>Inhale</strong>: 22 (8.9%) [6]</td>
</tr>
<tr>
<td>Smoke, smokes, smokies, smoking</td>
</tr>
<tr>
<td>Vaporized</td>
</tr>
<tr>
<td><strong>Inject</strong>: 14 (5.7%) [6]</td>
</tr>
<tr>
<td>Inject, injection</td>
</tr>
<tr>
<td>Boot</td>
</tr>
<tr>
<td>IVDU, needles</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

For the element *Time Frame* in Table 5, information about a past period of time was represented in research measures by values in the group ‘Past [#] [time unit]’ including phrases such as ‘during the last 30 days’ and ‘past 30 days up to and including today.’ In contrast, a past time period was represented in the group ‘Past’ containing more generalized values such as ‘former,’ ‘past,’ and ‘prior,’ in addition to the group ‘[#] [time unit] ago’ with more precise values including ‘10 years ago’ and ‘1 week ago’ within clinical text.

<table>
<thead>
<tr>
<th>Table 5: Distribution of values for <em>Time Frame</em> element. (# of values; (Frequency); [# unique values])</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical Text</strong></td>
</tr>
<tr>
<td>90 Total Values; 48 Unique Values; 9 Groups</td>
</tr>
<tr>
<td><strong>Past</strong>: 30 (33.3%) [6]</td>
</tr>
<tr>
<td>Former</td>
</tr>
<tr>
<td>Past, prior</td>
</tr>
<tr>
<td>Prev, previous</td>
</tr>
<tr>
<td><strong>Current</strong>: 18 (20.0%) [5]</td>
</tr>
<tr>
<td>Current, currently</td>
</tr>
<tr>
<td>Now, present</td>
</tr>
<tr>
<td>This time</td>
</tr>
<tr>
<td>[#] [time unit] ago: 9 (10.0%) [9]</td>
</tr>
<tr>
<td>Ten years ago</td>
</tr>
<tr>
<td>2 weeks ago</td>
</tr>
<tr>
<td>1 month ago</td>
</tr>
</tbody>
</table>

In Table 6, the element *Amount* was represented by the same four most frequently observed groups in both clinical text and research measures. The group ‘[#] [amount unit]’ contained the most values for measures at a frequency of 23.4% of 94 total values and 30.6% of 85 total values for clinical text, encompassing 22 and 23 unique values.
respectively. For both sources, the group ‘Other’ contained vague values for Amount including ‘any,’ ‘other amounts,’ ‘too much,’ ‘some,’ and ‘little’.

Table 6: Distribution of values for Amount element. (# of values; (Frequency); [# unique values])

<table>
<thead>
<tr>
<th>Clinical Text</th>
<th>Research Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>85 Total Values; 48 Unique Values; 9 Groups</td>
<td>93 Total Values; 32 Unique Values; 9 Groups</td>
</tr>
<tr>
<td>[#] [amount unit]: 26 (30.6%) [23]</td>
<td>[#] [amount unit]: 22 (23.4%) [22]</td>
</tr>
<tr>
<td>1 bowl</td>
<td>/_pills</td>
</tr>
<tr>
<td>10 bags</td>
<td>/_grams</td>
</tr>
<tr>
<td>2 cans</td>
<td>/_ampoules</td>
</tr>
<tr>
<td>None: 17 (20.0%) [2]</td>
<td>None: 17 (17.2%) [1]</td>
</tr>
<tr>
<td>None, nothing</td>
<td>Exact number: e.g., 0, 1, 2</td>
</tr>
<tr>
<td>[#]: 16 (18.8%) [11]</td>
<td>None: 13 (13.8%) [1]</td>
</tr>
<tr>
<td>1, 1-2, 6-7</td>
<td>None</td>
</tr>
<tr>
<td>one, x1</td>
<td>Other: 21 (23.5%) [4]</td>
</tr>
<tr>
<td>Other: 41 (37.4%) [9]</td>
<td>Any</td>
</tr>
<tr>
<td>Any, any significant, significant</td>
<td>Estimates</td>
</tr>
<tr>
<td>Some</td>
<td>Other amounts</td>
</tr>
<tr>
<td>Little</td>
<td>Too much</td>
</tr>
</tbody>
</table>

Discussion

Social and behavioral risk factors such as nicotine or alcohol use are known to impact health and are often documented in clinical settings. However, the collection of drug use information remains complex and often disregarded. Documenting drug use patterns and behavior in the EHR could support research and public health policy; aid physicians and other clinicians in identifying drug misuse, abuse, and dependence; and, highlight risk factors for comorbid conditions that could lead to enhanced patient care through prevention, intervention, and cost-effective, targeted treatment. The findings of this study provide insights into the current scope of the collection of drug use information in clinical and research settings, and reflect the wide variation in the breadth and depth of drug use content within standards, clinical text, and research measures. These findings may be used to inform development of a comprehensive drug use model for guiding improved data collection and use in electronic health sources (e.g., EHRs and surveys), which could be further enhanced by understanding the requirements of different end users and use cases.

Numerous sources of measures related to social and behavioral domains were identified in this study, including caDSR and PhenX Toolkit that were examined. Additional resources were also identified (e.g., GEM) that may be further explored in future work. Of the sources examined, PhenX measures contained the most robust drug use content since some measures contain multiple questions and are supplemented with additional information including lists of drug use methods, amounts, and drug types. By contrast, drug use measures identified in the caDSR using the CDE Browser were represented by a single question with additional information statements for most measures and were from resources including the NCI cancer Biomedical Informatics Grid (caBIG; seven measures), Lombardi Cancer Center (LCC; two measures), National Institute on Drug Abuse (NIDA; 26 measures), National Institute of Dental and Craniofacial Research (NIDCR; one measure), and NCI Programs of Research Excellence (SPOREs; four measures). The challenge of identifying drug use-specific measures across the multitude of resources highlights the need to develop a set of standard measures for this domain and supports the recent focus on adopting standard social and behavioral measures in the EHR.

Collectively, the analysis of standards, clinical text, and research measures revealed a wide-ranging set of drug use values that were represented by 27 elements. The Temporal element was further categorized into a total of nine elements. Start Date and Quit Age were predominantly represented in standards, whereas Start Age and Quit Date were found across the source types. The elements Duration of Use, Duration Since Quit, Duration Since Time Point, Time Frame, and Time Point appeared in clinical text and research measures, and overall, the values for each element were more precisely represented in clinical text. For example, clinical text contained Time Point values such as ‘Saturday,’ ‘15 years ago,’ ‘2006,’ and ‘July 4th’ while ‘first time,’ ‘last time,’ and ‘most recent time’ were found in the measures. These value disparities highlight the inherent difference in purpose between the two sources: clinical text reflects a more factual account of patient history, whereas measures are a clinically investigative tool seeking information using a generalized set of questions. Of note, reference to the age a person used a drug (e.g.,
‘age of 13-14’) appeared only within clinical text and was annotated as a value within Time Frame; however, in future work the element Use Age could be added to represent these values most accurately.

On occasion, the analysis of values presented an organizational challenge. Complex values containing multiple elements were observed in some measures. For example, within the element Amount, the value ‘#’ was further defined by the unit of measure including ‘bags,’ ‘buttons,’ ‘capsules,’ ‘hits,’ and ‘rocks,’ among other units which are nonstandard but have meaning in the drug use community. This finding thereby demonstrates the need for additional delineation of Amount values to include a numerical value as well as a unit of measure. Type values also presented a challenge since they were found in clinical text and research measures in the form of a slang or street name, or common, chemical, or pharmaceutical name. Interestingly, ‘caffeine’ was the second most common Type value in clinical text, motivating the exploration of a separate Caffeine model as a possible next step guided by existing standards such as the ‘Caffeine Consumption’ archetype in openEHR. Groupings for Type were designated either by a single value such as ‘party pills’ or by creating a group for words or phrases clearly representing one concept such as the group ‘Methamphetamine’ which contains the values ‘crank,’ ‘crystal meth,’ ‘meth,’ ‘desoxyn,’ and ‘fumes of crystal meth.’ Next steps for standardization and categorization of Type values may include exploring existing drug terminologies such as RxNorm from the National Library of Medicine44,45 that provides normalized names for clinical drugs, information and categorization of drugs from NIDA46, and the Alcohol and Other Drug (AOD) Thesaurus from the National Institute on Alcohol Abuse and Alcoholism (NIAAA)47. The availability of a comprehensive resource that hierarchically organizes drug types, variant names, and groupings could be valuable for providing flexibility to support different use cases (e.g., selecting a specific drug type versus a more general drug category).

The findings from this study represent a preliminary synopsis of drug use information identified within diverse sources with future goals involving the creation of a formal representation of an integrated drug use model. To that end, a next step could involve retrieving existing codes associated with measures identified in this study. For example, the PhenX measure ‘Substances – 30 day frequency’ correlates to the Logical Observation Identifiers Names and Codes (LOINC) concept with the name ‘PhenX - substance – 30 days frequency protocol,’ whereas the measure ‘Substance Abuse Illicit Substance Cocaine Personal Medical History Frequency’ from the CDE Browser contains the concepts ‘illicit substance,’ ‘cocaine,’ and ‘personal medical history’ with corresponding NCI Thesaurus codes. In addition, drug use values could be mapped to standardized terminologies and coding systems such as SNOMED-CT48 49 and the Unified Medical Language System (UMLS)50. Subsequent model development could also include alignment of these model components with national and international modeling initiatives such as the Clinical Information Modeling Initiative (CIMI)51, with the focus of generating a semantically interoperable drug use model.

Conclusion

Recent focus on promoting the collection of behavioral and social information in the electronic health record has highlighted the needs and challenges in developing standardized measures for these domains, including drug use. This study provides a broad perspective on the current representation of drug use information as reflected by standards, in documentation from clinical settings, and within research measures. The findings further provide a foundation for next steps in the development of a comprehensive drug use model that might be used to support research, clinical, and public health applications.

Acknowledgments

The authors thank Tamara Winden for discussions related to the annotation guidelines; Elizabeth Lindemann for contributing to the annotation of clinical notes; and, Yan Wang for supporting aspects of the annotation process. The clinical notes were obtained with permission from MTSamples (https://www.mtsamples.com). Research reported in this manuscript was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011364. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


Similarity-Based Recommendation of New Concepts to a Terminology

*Praveen Chandar, PhD, *Anil Yaman, MS, Julia Hoxha, PhD, Zhe He, PhD, Chunhua Weng, PhD

Department of Biomedical Informatics, Columbia University, New York, NY USA

(*equal contribution first-authors)

Abstract

Terminologies can suffer from poor concept coverage due to delays in addition of new concepts. This study tests a similarity-based approach to recommending concepts from a text corpus to a terminology. Our approach involves extraction of candidate concepts from a given text corpus, which are represented using a set of features. The model learns the important features to characterize a concept and recommends new concepts to a terminology. Further, we propose a cost-effective evaluation methodology to estimate the effectiveness of terminology enrichment methods. To test our methodology, we use the clinical trial eligibility criteria free-text as an example text corpus to recommend concepts for SNOMED CT. We computed precision at various rank intervals to measure the performance of the methods. Results indicate that our automated algorithm is an effective method for concept recommendation.

Introduction

Over years, researchers have stressed on the importance and usefulness of developing domain-specific terminologies in biomedical informatics. Well-curated terminologies provide cohesive and structured domain knowledge to clinical information systems, such as clinical decision support systems and Electronic Health Record systems, further facilitating the interoperability among them. Natural Language Processing (NLP) applications in biomedical informatics also largely rely on controlled terminologies and ontologies. Concept coverage is an important aspect of Cimino’s desiderata for terminologies (1). Good coverage requires timely and continuous concepts updating with terminologies to keep up with the non-stop growth of domain knowledge and the evolution of professional vocabularies. However, developing and maintaining a high-coverage terminology is a complex and expensive task that often requires constant laborious curation. Such manual effort involves domain experts who are required to read as much domain-specific literature as possible in order to identify meaningful concepts to insert into a terminology.

Due to limited resources and the subjectivity of domain experts in the manual curation process, terminologies vary in their coverage of domain concepts. As such, they are often criticized for insufficient coverage of concepts and concept relations such as synonyms (2). Recently, informatics tools are being used in order to ease the load on curators. BioPortal developed by, Nation Center for Biomedical Ontology (NCBO) (3) and Collaborative Protégé by Stanford (4) are some of the recent developments that facilitate continuous development of terminologies from geographically distributed locations. These tools facilitate continuous interaction between users and developers with the intent to improve terminology development through constructive feedback.

To complement the top-down approach for terminology development driven by domain experts, symbolic and statistical ontology learning methods have been proposed for identifying important concepts from human-generated texts. Hearst introduced a symbolic method to discover hyponyms within Lexical-Syntactic Pattern (LSPs) extracted from text (5). Liu et al. (6) further applied Hearst’s LSP method to identify clinically meaningful concepts and relationships from medical documents. They used regular expressions over Part-of-Speech (POS) tags to extract LSPs in medical documents and asked human annotators to identify meaningful concepts in these patterns. Church and Hanks proposed a statistical method for estimating word association norms to identify useful concepts (7). A symbolic-and-statistical hybrid approach was proposed to identify meaningful concepts by computing the similarity between unrecognized concepts and existing concepts in controlled terminologies (8). Recently, He et al. (9) proposed a structural method leveraging UMLS’s native term mapping and hierarchical structure to identify potentially missing concepts for a source terminology.

Even though the aforementioned approaches have achieved promising results, they all require human experts to suggest and review the concepts. In this work, we present a purely computational method for recommending new concepts with a minimum of human intervention. Unlike above mentions approaches, our method does not rely on domain heuristics or manually crafted rules; it characterizes the concepts using a set of features, which in turn is used to recommend new concepts to a source terminology. The contribution of this paper is two-fold:

(1) We propose a novel, unsupervised approach for prediction and recommendation of phrases as candidate atoms in an existing terminology based on concept similarities; and
(2) We present a cost-effective evaluation methodology to estimate the performance of the terminology enrichment method. Through experimental evaluation on large datasets, we show the accuracy of our recommendation approach, and demonstrate its potential to improve the scalability and throughput of terminology enrichment.

Methods

Glossary and Design Rationale

Before describing our methodology, we clarify the definition of concept, atom, and n-gram, which are key to understanding our approach.

- **Concept** is the fundamental unit of meaning in ontology. According to the semiotic triangle theory (10), each object in the world has one and only one concept describing it, but the concept may be associated with multiple terms, which are also called atoms. Concepts are assigned a unique identifier, such as the unique reference as an identifier (e.g., Concept Unique Identifier in the Unified Medical Language System (UMLS) that can be used to unambiguously identify a concept. Each concept is associated with at least one or more lexical variants or natural language text strings, which is often referred to as atoms in the UMLS.

- **Atom** is the smallest unit of meaning contributed by a source. Atoms are phrases (i.e. sequence of terms) that represent a distinguished meaning. A unique identifier is often assigned to an atom, such as the atomic unique identifier (AUI) in the UMLS. Atoms are symbols representing a concept, which contain one or more atoms.

- **N-Gram** is a contiguous sequence of n terms found in text. For example, *lymph node metastasis* is a 3-gram (or tri-gram). An n-gram deemed meaningful by a terminology curator becomes an atom.

The goal of our study is to identify in a collection of written texts of a particular domain, the n-grams that are highly probable of constituting an atom in an existing terminology; and to recommend them as candidate atoms to the curators. It is the decision of the curators to enrich the terminology accordingly by approving the candidates as atoms to an existing concept or a new concept in the terminology.

Our proposed method requires two components as input: seed terminology and a text corpus. We explain each of these components, our proposed approach to recommend new concepts and the evaluation strategy below.

Dataset

*Seed Terminology* is the terminology to be enriched using the text corpus. It must have at least two components: a set of concepts and a set of lexical variants or terms for each concept. Following de Keizer’s definition (11), a concept is a “cognitive construct of objects” that consists of one or more terms, which are also referred as atoms. For example, a concept “breast cancer” might have the following atoms: “Malignant neoplasm of breast”, “breast cancer”, etc. In the UMLS, atoms are often assigned an atomic unique identifier (AUI).

*Text Corpus:* A large set of unstructured text on a specific domain from which concepts are to be extracted. Text documents and literature are valuable sources for identifying new concepts on a domain. The written text must ideally be a representative sample of the knowledge prevalent in the domain. It is important that the text exists as free-text as our approach relies on occurrence frequency as well as linguistic properties of a concept in the text. The goal in this work is to identify potential atoms from the text to enrich an existing terminology.

![System outline illustrating various modules in our proposed approach.](image)

**Figure 1 – System outline illustrating various modules in our proposed approach.**
Approach

Our proposed approach takes two data sources as input: seed terminology and text corpora, and outputs a ranked list of n-grams that are most likely to be added to the seed terminology as atoms. The fundamental premise of our approach is that the characteristics of n-grams that are already atoms in the seed terminology can be automatically learned and used to recommend new atoms that can be added to the seed terminology.

Figure 1 provides an outline of our approach. For a given text corpus, the text is preprocessed by identifying set of sentences. Then, a set of n-grams is extracted from each sentence. The extracted n-grams are aggregated and each unique n-gram is characterized using a set of features. We use a set of morphological, contextual, and syntactic features to capture the characteristics of an n-gram that represents a concept. Of all n-grams identified in the text corpus, a subset of them can be matched to atoms in a seed terminology. We leverage this information to learn a model that characterizes an atom using a set of features. Considering each atom identified in the text corpus as a candidate atom, the model is used to score these candidate atoms indicating the likelihood of them being suitable to be an atom in the seed terminology. A summary of various steps involved are listed below:

1. Preprocessing the text corpus;
2. Identifying n-grams from text and extraction of features for each n-gram;
3. Learning the characteristics of those n-grams that belong to the seed terminology;
4. Leveraging the learnt characteristics to recommend new atoms to the terminology;

Step 1: Pre-processing: The first step is to segment the free-text from the text corpus into sentences. We relied on OpenNLP (12) to perform sentence segmentation. Each sentence was tokenized and n-grams were extracted by iterating over the tokens (up to 5-grams are extracted). The extracted n-grams were filtered if the n-gram begins with a number token or a stop-word, or if the n-gram ends with a stop word. We used a Standard English stop-word list in our experiments.

Step 2: Feature Extraction: The problem of identifying meaningful atoms in free text is similar to the named entity recognition (NER) problem (13). Inspired by the research in NER, we engineered a set contextual and syntactic features to represent the n-grams extracted from the text corpus (13, 14). The feature representation of n-gram is not specific to any terminology or text corpora. This is an important property of our method that enables our method to be used across other domains. The engineered features include:

- Capitalization Features (CAP): the capitalization information of the first and all letters, and the first letters of each tokens of the current n-gram;
- Syntactic Features (POS): the part-of-speech (POS) tags of the current n-gram;
- Contextual Features (CONTEXT): the POS-tags, tokens and the prefixes/suffixes of the tokens that co-occur with the current n-gram within a window of 10 tokens;

![Figure 2 – Feature extraction example for the unigram “chemotherapy”. (a) Capitalization Features, (b) Syntactic Features, and (c, d, and e) Contextual Features.](image-url)
We illustrate the feature extraction process with an example in Figure 2, the example shows two different sentences containing “chemotherapy”, which is the unigram under consideration. The feature extraction process starts with the tokenized and POS-tagged sentences. In Figure 2, (a) represents the capitalization features; (b) illustrates the syntactic features of the unigram; and (c, d and e) are the contextual features that include frequencies of POS-tags, tokens, and prefixes/suffixes of the surrounding tokens within a windows size of 10. As suggested in the literature, the length of the prefixes and suffixes was limited to 3 characters. The feature vector of the unigram was obtained by concatenating of the features shown in (a, b, c, d and e).

Step 3: Learning Module: Once the n-grams were characterized by a set of features, the next step was to build a model that accurately represents n-grams that were the concepts belonging to the seed terminology using our feature representation. Thus, we split the n-grams identified into two sets: n-grams that were concepts in the seed terminology (Concept Set) and n-grams that were not in the seed terminology (Candidate Atom Set). Note that the concept set here was automatically generated using the seed terminology and the text corpus.

The n-grams in the candidate atom set are represented using a set of features, and clustered using the K-means algorithm as shown in Figure 3. The rationale behind clustering is that the groupings would represent different characteristics of an n-gram that are atoms in the seed terminology. We hypothesize that the clustering categorizes the characteristics of the atoms in a seed terminology, thereby improving recommendation of candidate atoms. Determining the right number of clusters is not obvious for a given dataset, several methods exist for the finding the optimal k in K-means (15, 16). We prune those clusters that contain less than 1% of the total n-gram. The pruning was performed to eliminate less representative clusters that could introduce noise into our model.

Centroid linkage function: the minimum Euclidean distance between the feature vector of an n-gram and the cluster centroids.

\[
\text{score}_i = \min\left( \text{dist}(x_i, c_j) \right), \quad j \in \{1, 2, \ldots, C\}
\]

Where, \( i \) is the \( i^{th} \) n-gram in the set; \( x_i \) is the feature vector of the \( i^{th} \) n-gram; \( C \) is the cluster count; \( c_j \) feature vector of the centroid in \( j^{th} \) cluster, and \( \text{dist}(x_i, c_j) \) is the function that measures the Euclidean distance between two vectors.

Number of Clusters: The value of \( k \) in k-means is determined using a cross-validated likelihood criterion (16). The method follows a cross-validated clustering approach to provide insights about cluster structure. The concept set (CS) was split into 10 folds. For each fold, samples from the training folds were clustered using k-means with different \( k \) values.

The sum of average differences between the samples in the test set and the nearest cluster centroid were computed. The computed sum is called the average minimum distance. Figure 4 shows the average minimum distance for
various $k$ values. It can be seen that the minimum average distance stabilizes for $k$ values between 30 and 40, which can be used for an indication of the optimum number of the clusters.

![Graph showing the average minimum distance scores for various $k$-values.](image)

**Figure 4 - The average minimum distance scores for various $k$-values.**

**Evaluation**

In this section, we describe the evaluation measures and the details of the dataset used in our study to evaluate our proposed approaches.

**Evaluation Measure:** We use precision at various rank levels ($r$) to evaluate our atom recommendation algorithm. Precision at rank $r$ is commonly used for various ranking problems to measure the effectiveness. Precision at rank $r$ is the fraction of retrieved n-grams that are indeed meaningful atoms.

$$\text{Precision}@r = \frac{\# \text{of n-grams identified as meaningful atoms}}{n}$$

Where, $k$ is the rank cut-off. We compute precision at $r = 1000, 5000$ and $10,000$, as it is reasonable to assume that a terminology curator enriching a terminology such as SNOMED CT (containing about 300,000 active concepts) would be willing to manually go through at least 5000 n-grams. Since, the task at hand is to re-rank the atoms in the candidate atom set (CAS), traditional ranking measures such as recall and F-measure are not informative.

**Domain-Specific Text Corpus:** We use the ClinicalTrials.gov dataset that is a rich source of information containing a set of eligibility criteria used by clinical researchers (17). The dataset was created by the National Library of Medicine and is publicly available. Each clinical trial consists of various structured fields such as study title, sponsor, etc., and a free text eligibility criteria field that contains the criteria for inclusion and exclusion of a participant. The eligibility criteria define various patient characteristics that make a person eligible or ineligible for a research study. We make use of the eligibility criteria free text field for the purpose of this study. We downloaded 181,356 (as of February 2015) clinical trial summaries for our experiments.

**Seed Terminology:** For the purpose of this experiment, we assume that SNOMED CT is the only available terminology at hand so that we do not consider the possibility of one term belonging to multiple terminologies simultaneously. SNOMED CT is one of most comprehensive clinical publicly available clinical terminology and is maintained and distributed by International Health Terminology Standards Development Organization (IHTSDO). We obtained the latest version of SNOMED CT available as part of the UMLS Metathesaurus (version 2014AA) from the NLM website. Our algorithm intends to extend SNOMED CT terminology with atoms relevant to the clinical domain using free-text clinical trial eligibility criteria.

**Labeled Data:** In order to evaluate our proposed methodology, we require a set of n-grams with positive and negative labels, i.e., positive n-grams are meaningful atoms whereas negative refers to a meaningfulness n-gram. As human annotations are extremely expensive and time consuming, we built our labeled data using various terminologies that are part of the UMLS, which were curated by experts from different domains. All n-grams identified in the ClinicalTrials.gov dataset that were part of a terminology (other than SNOMED) were labeled positive and the rest were labeled negative. It is certainly possible that a negatively labeled n-gram could still be meaningful, UMLS terminologies are not complete and there may still be meaningful n-grams not part of the UMLS. Nevertheless, the simulated labeled data provides a cost-effective, quick, and reasonable lower-bound performance estimate of our proposed methods.
Results

Dataset Statistics

We used the free-text clinical trial eligibility criteria to extract all possible n-grams. The n-grams were preprocessed to remove stop-words and numbers as described in the pre-processing step. We extracted up to 5-grams that occurred at least 5 times in the corpus to filter out misspelled or noisy n-grams. The pre-processed n-grams were categorized into concept set (CS) and candidate atom set (CAS). An n-gram that matched against an atom in SNOMED belonged to the CS and all other n-grams belonged to CAS.

Table 1 shows the total number of unique n-grams extracted. Only 5% of those n-grams matched against an atom in the SNOMED. Of the remaining 429,465 n-grams, about 6% (25,291 n-grams) matched against an atom in the UMLS (other than SNOMED). The 25,291 n-grams constitute our positive labels and are used only for evaluation purpose. The remaining 404,174 n-grams belong to the candidate atoms set (CAS); the n-grams in CAS are to be ranked by our recommendation algorithm.

<table>
<thead>
<tr>
<th>Total Unique n-grams (1 to 5 tokens)</th>
<th>449,978</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Set (CS)</td>
<td>Concept Set (SNOMED CT)</td>
</tr>
<tr>
<td>Candidate Atoms Set (CAS)</td>
<td>Labeled Atoms (UMLS–SNOMED CT)</td>
</tr>
<tr>
<td></td>
<td>Number of Unmatched N-grams</td>
</tr>
</tbody>
</table>

Lower Bound Precision Scores

The candidate atoms in CAS were ranked by our algorithm, which promoted n-grams that are relevant to the seed terminology higher and suppressed meaningless n-grams. An n-gram in CAS that matched against an atom in UMLS (other than SNOMED CT) was considered as true positive. Those n-grams that do not match are considered as true negatives. We computed precision at various rank intervals (1000, 5000, 10000 and 20000), to estimate the effectiveness of our approach.

As described in the methods section, the choice of features used to represent n-grams influences the performance of the recommendation algorithm. Therefore, we experiment with different sets of features. We experimented with different feature sets and their combination of the capitalization (CAP), syntactic (POS) and contextual features (CONTEXT). In order to observe the effect of the syntactic information, we merged all the syntactic information under the feature set named POS (including the syntactic information involved in the contextual feature set).

In addition, we included two baseline methods for the comparison: a random ranking and frequency-based. The random ranking baseline was obtained by randomly sampling 20,000 n-grams from the candidate atom set (CAS). The frequency-based ranking was obtained by sorting n-grams in the candidate atom set (CAS) by frequency of their occurrence in the text corpora.

Table 2 provides the accuracy scores for various methods at various rank intervals r. The results show that our algorithm performs considerably better than a frequency-based baseline. Note that, especially for larger r values the difference between our recommendation algorithm and the frequency baseline is more apparent. It is observed that the inclusion of the POS features further increases the precision by 5%. The best results were obtained using all the features (CAPS+POS+CONTEXT) that represent an atom.

Table 2 – Lower Bound Precision scores at different rank intervals for various methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Lower Bound Precision @ r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=1000</td>
</tr>
<tr>
<td>CAPS+POS+CONTEXT</td>
<td>0.743</td>
</tr>
<tr>
<td>CAPS+POS</td>
<td>0.695</td>
</tr>
<tr>
<td>CAPS+CONTEXT</td>
<td>0.550</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.529</td>
</tr>
<tr>
<td>Random</td>
<td>0.047</td>
</tr>
</tbody>
</table>

Table 2 provides the accuracy scores for various methods at various rank intervals r. The results show that our algorithm performs considerably better than a frequency-based baseline. Note that, especially for larger r values the difference between our recommendation algorithm and the frequency baseline is more apparent. It is observed that the inclusion of the POS features further increases the precision by 5%. The best results were obtained using all the features (CAPS+POS+CONTEXT) that represent an atom.
Top-10 n-grams
Table 3 shows the top ranked atoms categorized by 1, 2, 3, 4, and 5 grams. The atoms that belong to a terminology in the UMLS are italicized and the rest are highlighted in grey. It is interesting to note that atoms such as “history of sleep apnea”, “hormonal treatment” are meaningful but are not part of any existing UMLS terminology.

Examining the top ranked n-grams lead to a few interesting observations. We found that there existed meaningful n-grams that did not match against the UMLS in the top ranked list. Thus, we decided to investigate this further by estimating the number of meaningful n-grams (candidate atoms) that are not part of the UMLS. We used the help of a physician to determine if the candidate atoms are meaningful concepts. The physician was provided a list of 100 n-grams that were not matched against a UMLS terminology for manual review.

The top 100 n-grams were obtained using the algorithm with the best precision score (as shown in in Table 2); and 37 of the n-grams in this were labeled as meaningful. The annotations clearly indicate that there exist several n-grams that are currently not part of the UMLS. The annotation effort was preliminary step and we plan to perform a comprehensive evaluation in the future.

Table 3 – Top 10 n-grams (n ∈ [1-5]) ranked by our recommendation algorithm. N-grams that matched to UMLS concepts are italicized, and n-grams that did not match are highlighted in gray.

<table>
<thead>
<tr>
<th>UNIGRAMS</th>
<th>BIGRAMS</th>
<th>TRIGRAMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>facial</td>
<td>immune suppression</td>
<td>collagen vascular disease</td>
</tr>
<tr>
<td>opioids</td>
<td>hormonal treatment</td>
<td>estrogen replacement therapy</td>
</tr>
<tr>
<td>activation</td>
<td>medical treatment</td>
<td>bone marrow suppression</td>
</tr>
<tr>
<td>treatment</td>
<td>cell therapy</td>
<td>substance abuse treatment</td>
</tr>
<tr>
<td>content</td>
<td>immunosuppressive treatment</td>
<td>lymph node metastasis</td>
</tr>
<tr>
<td>strength</td>
<td>pain treatment</td>
<td>lymph node involvement</td>
</tr>
<tr>
<td>coil</td>
<td>pharmacological treatment</td>
<td>significant cognitive impairment</td>
</tr>
<tr>
<td>genetic</td>
<td>functional impairment</td>
<td>tyrosine kinase inhibitor</td>
</tr>
<tr>
<td>resistance</td>
<td>vertebral fracture</td>
<td>diagnosis of cancer</td>
</tr>
<tr>
<td>titration</td>
<td>cognitive deficit</td>
<td>severe liver disease</td>
</tr>
<tr>
<td>4-GRAMS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>central nervous system involvement</td>
<td>severe chronic obstructive lung disease</td>
<td></td>
</tr>
<tr>
<td>stable coronary artery disease</td>
<td>history of congestive heart failure</td>
<td></td>
</tr>
<tr>
<td>central nervous system disease</td>
<td>history of coronary artery disease</td>
<td></td>
</tr>
<tr>
<td>decompensated congestive heart failure</td>
<td>history of ischemic heart disease</td>
<td></td>
</tr>
<tr>
<td>active urinary tract infection</td>
<td>autologous hematopoietic stem cell transplantation</td>
<td></td>
</tr>
<tr>
<td>diagnosis of breast cancer</td>
<td>history of traumatic brain injury</td>
<td></td>
</tr>
<tr>
<td>multivessel coronary artery disease</td>
<td>evidence of coronary artery disease</td>
<td></td>
</tr>
<tr>
<td>history of seizure disorder</td>
<td>low white blood cell count</td>
<td></td>
</tr>
<tr>
<td>history of sleep apnea</td>
<td>radiographic evidence of disease progression</td>
<td></td>
</tr>
<tr>
<td>nonalcoholic fatty liver disease</td>
<td>history of substance use disorder</td>
<td></td>
</tr>
</tbody>
</table>

New Concepts vs. Synonyms
The precision at rank interval is computed by dividing the number of n-gram identified as an atom of the concept by rank interval. The number of n-grams can be categorized into two groups: n-grams that would be added to the terminology as a new concept from the atoms that will be added as synonyms to an existing concept. This enables us to separate atoms of new concepts from atoms of existing concepts in the terminology. The analysis on our results revealed that on average at least 45% of the n-grams were atoms of existing concepts (often referred to as synonyms in the literature) and the remaining were atoms that belong to new concepts to be added to the seed terminology.

Discussion

Cluster Analysis
In the learning module, the n-grams in the concept set (i.e. n-grams that are atoms in SNOMED CT) are clustered based on their distance between the feature vectors. We hypothesized that the clustering would group similar atoms into the same cluster. We analyzed each cluster by manually reviewing 100 n-grams closest to the cluster centroids. The intention was to identify similarities between n-grams in the cluster or manual label.

Table 4 – Representative n-grams from the n-gram clusters

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>mexico</td>
<td>extent of disease</td>
<td>pta</td>
<td>repeated</td>
</tr>
<tr>
<td>illinois</td>
<td>quality of life</td>
<td>mm</td>
<td>improved</td>
</tr>
<tr>
<td>canada</td>
<td>range of motion</td>
<td>pd</td>
<td>restricted</td>
</tr>
<tr>
<td>europe</td>
<td>lactose intolerance</td>
<td>dt</td>
<td>complicated</td>
</tr>
<tr>
<td>denmark</td>
<td>history of anemia</td>
<td>cvd</td>
<td>involved</td>
</tr>
<tr>
<td>kenya</td>
<td>history of surgery</td>
<td>hctz</td>
<td>healed</td>
</tr>
<tr>
<td>africa</td>
<td>pain at rest</td>
<td>ctx</td>
<td>provoked</td>
</tr>
<tr>
<td>vermont</td>
<td>history of depression</td>
<td>ttp</td>
<td>stabilized</td>
</tr>
<tr>
<td>indiana</td>
<td>level of consciousness</td>
<td>nsaid</td>
<td>established</td>
</tr>
</tbody>
</table>

Table 4 illustrates a set of 10 n-grams for four clusters that are closest to their centroids. Interestingly, Cluster 1 tends to consist of atoms relating to geographic locations or place names. Cluster 3 contains abbreviations. The atoms in Cluster 2 contain atoms following a specific pattern, this is somewhat expected due to the fact that the atoms are represented using syntactic features. Cluster 4 groups adjectives that are prevalent in the clinical domain.

Further, we tried to align the clusters with the semantic hierarchies present in SNOMED CT. The concepts in SNOMED CT are organized into 19 acyclic taxonomic (is-a) hierarchies. We observed the distribution of the semantic hierarchies in each cluster. For example, 48% of the concepts found within a cluster belonged to the hierarchy “Staging & Scales”, 72% of the concepts in another cluster belonged to “Situation with Explicit Context”. However, only a weak correlation was observed between the clusters and SNOMED CT hierarchies.

**Modifier Extraction**

Modifiers are terms or phrases that provide additional meaning to a concept. For instance, in the phrase *history of heart disease*, the prefix *history of* modifies the meaning of the concept *heart disease*. Modifiers play an important role in tasks such as cohort identification. A researcher looking for cohort of patients with “history of heart disease” might possibly find little use with patients with “active heart disease”. As an extension to this work, our intention is to identify candidate modifier that enrich the identified concepts.

Table 5 – Candidate modifiers identified from the top ranked n-grams.

<table>
<thead>
<tr>
<th>history of</th>
<th>severe</th>
<th>significant</th>
<th>active</th>
<th>previous</th>
</tr>
</thead>
<tbody>
<tr>
<td>current</td>
<td>evidence of</td>
<td>diagnosis of</td>
<td>treatment with</td>
<td>prior</td>
</tr>
</tbody>
</table>

We manually analyzed the top ranked n-gram and found that the 3-, 4- and 5-grams tend to contain a modifier followed by a concept. Following the manual analysis, we wanted to identify those n-grams that matched the pattern; prefix followed by UMLS atom. Using regular expressions we filtered all n-grams that matched the <prefix, atom> pattern. We found 14,000 n-grams that contain a prefix followed by a UMLS atom, i.e., partially matched against an atom in UMLS. Further, we ranked the prefix by its frequency; hypothesizing that commonly used modifiers would be ranked higher.

Table 5 shows the top 10 candidate modifiers extracted using this approach. The results are promising and we plan to investigate further in this direction.
**Post & Pre Coordination**

SNOMED Clinical Terms (SNOMED CT), as the most comprehensive multilingual clinical terminology, is being widely implemented as a standard within HITSDO member countries. By 2015, SNOMED CT will be the United States standard for encoding diagnoses, procedures, and vital signs in electronic health records (EHRs) under Stage 2 of Meaningful Use (18). Even though SNOMED CT provides rich conceptual content, researchers have advocated greater coverage of common problem statements with improved synonymy and conceptual content (19). A survey among the direct users of SNOMED CT reported that 23% and 17% of the respondents encountered missing concepts and missing synonyms, respectively (2).

Post-coordination of SNOMED CT allows its users to create new meaning by combining existing concepts, which can potentially enhance SNOMED CT’s conceptual coverage (20). In spite of post-coordination, researchers reported that some clinical statements with complex and rare clinical scenarios could not be encoded (21). Meanwhile, the same clinical meaning may be represented by different post-coordinated expressions, which hampers their interoperability among different authoring entities. In the setting of this paper, post-coordinated expressions may also pose difficulties in unifying and structuring clinical trial eligibility criteria.

To the best of our knowledge, there is no tool available for automating the creation of post-coordinated expressions. Therefore, they have to be modeled by a clinical expert manually. In this proof-of-concept study, we consider only pre-coordinated SNOMED CT concepts for training and prediction. Computational approaches to filter out suggested concepts that can be constructed with post-coordination need to be developed. Nevertheless, the suggested concepts were identified to be important in clinical trial eligibility criteria, thereby should be considered by SNOMED CT curators as pre-coordinated concepts.

**Limitations**

Since the recommendation algorithm relies on linguistic features of atoms extracted from the text corpus, the corpus must contain atoms already present in the seed terminology. Also, our method recommends atoms of a concept and does not provide any information about the relationship between the atoms and concepts. For instance, the algorithm would identify “sleep disordered breathing” and “breathing disorder during sleeping” as meaningful atoms although it provides no information w.r.t its relation with the concept Sleep Apnea. The problem of finding if an atom belongs to an existing or new concept is still an open research area that warrants more investigation. Another limitation is that, we rely on exact string matching between n-grams and atoms in a terminology, which could be error-prone. Although fuzzy matching of text is possible, we argue that an exact string matching provides a more accurate estimation.

**Future work**

In the future, we intend to conduct a larger-scale comprehensive evaluation study using domain experts to accurately estimate the performance of our methods. The manual evaluation process would require the experts to review each candidate atom to determine if the recommended atoms are meaningful or not. Also, we plan to compare our approach with previously proposed symbolic and statistical ontology learning methods (3, 6, 7). Another avenue of future work is to recommend sentences that are highly probable contain new atoms/concepts. Ranking sentences has an advantage, it provides context to the atoms that enables curators to make better decision about atom’s usefulness to a terminology. We also plan to experiment with different datasets from heterogeneous domains to prove the generalizability of our proposed approach.

**Conclusions**

This paper contributes a similarity-based approach for recommending candidate n-grams to terminology curators for consideration. The method characterizes the n-grams in a text corpus by using a feature set to learn important features of a concept in order to produce a ranked list of n-grams by decreasing order of meaningfulness. The ranked list of n-grams would enable quicker and easier identification of new concepts for an existing terminology. Another contribution of this paper is a cost-effective evaluation methodology to estimate the performance of the terminology enrichment method. Through careful experimental design, we simulate labeled data using various terminologies in UMLS to evaluate effectiveness of our proposed approach. We computed the precision score and observed that our method provides about 20% improvement over a frequency-based baseline. We have only scratched the surface of what is possible. There is ample opportunity for work within this framework such as scoring methods for ranking candidate atom, and post filtering of candidate atoms.
Acknowledgments

We would like to thank Dr. Kin Wah Fung from Lister Hill National Center for Biomedical Communications for the informative discussion. We would also like to thank Fendanda Polubriaginof M.D. for her help with manual annotations. This study was sponsored by the U.S. National Library of Medicine grant R01LM009886 (PI: Weng) and U.S. National Center for Advancing Translational Science grant UL1 TR000040 (PI: Ginsberg).

References

Mining and Visualizing Family History Associations in the
Electronic Health Record: A Case Study for Pediatric Asthma

Elizabeth S. Chen, PhD\textsuperscript{1,2}, Genevieve B. Melton, MD, PhD\textsuperscript{5,6},
Richard C. Wasserman, MD, MPH\textsuperscript{3,4}, Paul T. Rosenau, MD, MS\textsuperscript{3,4},
Diantha B. Howard, MS\textsuperscript{2}, Indra Neil Sarkar, PhD, MLIS\textsuperscript{1,2}
\textsuperscript{1}Center for Biomedical Informatics, Warren Alpert Medical School of Brown University,
Providence, RI; \textsuperscript{2}Center for Clinical and Translational Science and \textsuperscript{3}Department of
Pediatrics, University of Vermont, Burlington, VT; \textsuperscript{4}University of Vermont Children’s
Hospital, Burlington, VT; \textsuperscript{5}Institute for Health Informatics and \textsuperscript{6}Department of Surgery,
University of Minnesota, Minneapolis, MN

Abstract

Asthma is the most common chronic childhood disease and has seen increasing prevalence worldwide. While there
is existing evidence of familial and other risk factors for pediatric asthma, there is a need for further studies to
explore and understand interactions among these risk factors. The goal of this study was to develop an approach for
mining, visualizing, and evaluating association rules representing pairwise interactions among potential familial
risk factors based on information documented as part of a patient’s family history in the electronic health record. As
a case study, 10,260 structured family history entries for a cohort of 1,531 pediatric asthma patients were extracted
and analyzed to generate family history associations at different levels of granularity. The preliminary results
highlight the potential of this approach for validating known knowledge and suggesting opportunities for further
investigation that may contribute to improving prediction of asthma risk in children.

Introduction

Asthma is the leading chronic condition in children globally and has seen a dramatic increase in prevalence over the
last few decades\textsuperscript{1-3}. Several reports have described the need to better understand the effect of genetic, environmental,
and lifestyle factors and their interactions for improving pediatric asthma management\textsuperscript{4-7}. With respect to familial
risk factors, a number of studies have focused on exploring family history of asthma or atopic disease as a potential
predictor of asthma risk in children\textsuperscript{8-11}. These studies have primarily involved use of questionnaires and examined
the effect of asthma in first- and second-degree relatives (parents, siblings, and grandparents) and in multiple
relatives. One study described the low positive predictive value of a family history of asthma; however, it did
conclude that this could be a useful predictor for some preventive health efforts and further studies are needed to
understand the role of family history in risk assessment for pediatric asthma\textsuperscript{10}.

The electronic health record (EHR) offers a potentially valuable source for supporting such studies since information
pertaining to potential familial risk factors is increasingly collected as part of a patient’s family history in addition to
patient-specific diseases and conditions longitudinally over time\textsuperscript{12,13}. This information may be documented in
multiple places throughout the EHR in structured or free-text form, thus requiring the development of approaches to
extract, standardize, and integrate the information for subsequent analysis. To date, there have been some efforts
focused on improving documentation and use of family history in the EHR including a Stage 2 Meaningful Use
measure specifying structured data entry for one or more first-degree relatives in more than 20% of patients\textsuperscript{14},
specifications and models for representing family history\textsuperscript{15,16}, and development of natural language processing
(NLP) techniques for extracting family history from clinical notes\textsuperscript{17-20}.

Knowledge discovery and data mining approaches have the potential to transform EHR data into actionable disease
knowledge\textsuperscript{21-24}. Such approaches have been used to study a variety of disease relationships (e.g., disease-disease and
disease-drug) using both structured data (e.g., from problem and medication lists) and free-text within clinical notes
from the EHR\textsuperscript{25-29}. Association rule mining is a commonly used data mining technique for discovering “interesting”
relationships between items (e.g., family history) in large datasets (e.g., from the EHR) that could contribute to
generating hypotheses for further investigation\textsuperscript{30}. An association rule is represented as $X \Rightarrow Y$ (e.g., if patient has a
family history of $X$, patient also has a family history of $Y$) where $X$ is referred to as the antecedent or left-hand-side
(LHS) of the rule and $Y$ is referred to as the consequent or right-hand-side (RHS) of the rule. A number of
“interestingness” measures (e.g., support, confidence, lift, chi-square, and odds ratio) can be calculated to convey
the strength of a given rule\textsuperscript{31}. Typically, minimum thresholds are specified for support (e.g., proportion of patients
with a family history of $X$ and $Y$ relative to all patients) and confidence (e.g., proportion of patients with a family history of $X$ and $Y$ relative to patients with a family history of $X$) as a mechanism for constraining results.\textsuperscript{30}

The collection of family history data in EHR systems provides an opportunity for validating known associations and for potentially discovering new knowledge of interactions among familial risk factors and diseases. The objective of this study was to develop an approach for mining and visualizing family history associations in the EHR using open-source technologies. As a case study, the approach was evaluated for a particular condition, pediatric asthma.

**Methods**

Figure 1 provides an overview of the approach used in this study that is based on the processes for Knowledge Discovery in Databases\textsuperscript{32} and Disease Knowledge Discovery\textsuperscript{23}. The four major steps involved: (1) data selection to identify a cohort of pediatric asthma patients and extract associated family history entries from the EHR, (2) preprocessing and transformation to prepare the dataset for mining and visualization at different levels of granularity, (3) data mining to generate basic statistics and association rules, and (4) interpretation and evaluation to visualize and validate the family history associations. A combination of Ruby (2.0.0) and R (3.1.2), integrated using the RinRuby Ruby gem (2.0.3), were used for the various processing and analysis tasks.

**Data Selection**

At the University of Vermont Children’s Hospital, the Epic EHR (Epic Systems Corporation, Verona, WI) has been in use since 2009 and includes a module consisting of both structured and free-text fields for collecting family history information.\textsuperscript{33} Two of the structured fields were the focus of this study: (1) problem – selected from a list of 222 values such as “Asthma,” “Cancer,” or “*” for Other that can be further specified in a free-text comment field (not included in this study) and (2) relation – selected from a list of 22 values including “Father,” “Maternal Grandmother,” “Other” that can be further specified in the free-text comment field (not included in this study), or “Neg Hx” for indicating a negative family history of the selected problem (Table 1).

A dataset of structured family history entries was created for a cohort of pediatric asthma patients identified using the following criteria: (1) at least one encounter in 2014, (2) age 3 to <18 years old, and (3) ICD-9-CM code 493* as an encounter diagnosis in 2014 or on the problem list. The most recent set of family history entries (consisting of the problem and relation fields) associated with an encounter in 2014 were then obtained for each patient in the cohort.

**Table 1. Examples of Preprocessed and Transformed Family History Entries.**

<table>
<thead>
<tr>
<th>Problem</th>
<th>Relation</th>
<th>Comment</th>
<th>Side of Family</th>
<th>Degree of Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>Mother</td>
<td>as a child</td>
<td>maternal</td>
<td>first</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Paternal Grandfather</td>
<td></td>
<td>paternal</td>
<td>second</td>
</tr>
<tr>
<td>*</td>
<td>Brother</td>
<td>healthy</td>
<td>-</td>
<td>first</td>
</tr>
<tr>
<td>Cancer</td>
<td>Other</td>
<td>maternal great-aunt</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cystic Fibrosis</td>
<td>Neg Hx</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*light grey shading indicates excluded field or entry for this study; “*” indicates not applicable or cannot be inferred from structured relation field.
Preprocessing and Transformation

The full dataset was divided into two subsets for: (1) positive family history – entries associated with positive family history of a problem (i.e., any relation except for “Neg Hx”) and (2) negative family history – entries associated with negative family history of a problem (i.e., relation specified as “Neg Hx”). Within these subsets, entries with “*” selected as the problem were excluded since they do not specify a particular problem in structured form (but may include information in the free-text comment field).

For the positive family history subset, four datasets were created to enable subsequent analysis at different levels of granularity:

1. Family history of problem P – only includes the problem from entries for each patient (pos_problem dataset)
2. Family history of problem P in relation R – includes problem and relation from entries for each patient (pos_problem_relation dataset)
3. Family history of problem P in side of family S – includes problem and side of family (i.e., maternal or paternal) determined based on specified relation (pos_problem_side dataset) (Table 1, Column 4)
4. Family history of problem P in degree of relationship D – includes problem and degree of relationship (i.e., first or second) determined based on specified relation (pos_problem_degree dataset) (Table 1, Column 5)

For the negative family history subset, one dataset was created for negative family history of problem P that only included the problem from entries for each patient (neg_problem dataset).

The five resulting datasets were transformed into a common format in preparation for the subsequent data mining step. In this format, each row includes the patient identifier (e.g., random number) and information associated with a single entry separated by the “|” character. For example, “2|Asthma|” in the pos_problem dataset, “2|Asthma_Mother|” in the pos_problem_relation dataset, “2|Asthma_maternal|” in the pos_problem_side dataset, and “2|Asthma_first|” in the pos_problem_degree dataset.

Data Mining

For each dataset, basic statistics were calculated for ranking problems based on frequency and prevalence (pos_problem, pos_problem_relation, and neg_problem datasets only for the latter). Prevalence of a family history problem was determined relative to all family history entries associated with encounters in a particular year (2014 for this study) using a formula similar to term frequency-inverse document frequency (TF-IDF) that is often used in information retrieval to reflect importance of a term relative to a document in a collection:

\[
PREV(p_d) = \frac{\sum p_d}{N_d} \times \log \left( \frac{N_y}{\sum p_y} \right)
\]

where \( p_d \) represents patients with a family history of problem \( p \) in a cohort for a particular disease \( d \) (pediatric asthma), \( N_d \) is the total number of patients in the disease cohort, \( N_y \) is the number of patients with family history entries in year \( y \) (e.g., 2014), and \( p_y \) is the number of patients with a family history of problem \( p \) in that year.

Association rule mining was performed using the arules R package (1.1-6) that interfaces with a C implementation of the Apriori algorithm. With arules, minimum thresholds for support and confidence as well as maximum length of rules can be specified. Numerous other interestingness measures such as lift, chi-square (\( \chi^2 \)), and odds ratio can also be calculated for further filtering and ranking of the generated rules. Analysis of each dataset was performed using different combinations of minimum support values (0.0 to 0.1 in 0.01 increments) and minimum confidence values (0.0 to 1.0 in 0.1 increments). The maximum rule length was restricted to two in order to focus on pairwise associations. In addition to support and confidence, \( \chi^2 \) was also calculated and used to rank rules as this measure was found to outperform other measures in prior studies.

Interpretation and Evaluation

To facilitate interpretation, the arulesViz R package (1.0-0) was used that implements ten visualization techniques for exploring association rules generated by arules. Three types of visualizations were selected for viewing all rules or subsets of rules (i.e., top 50) in each dataset: (1) scatter plot that highlights the distribution of rules relative to specified measures (e.g., support, confidence, and \( \chi^2 \)), (2) graph-based that displays vertices representing items or itemsets and edges representing relationships in rules, and (3) grouped matrix-based that uses a balloon plot for
displaying rule antecedent groups as columns and rule consequents as rows. For the latter two visualizations, support and \( \chi^2 \) were specified as the measures for depicting rule strength based on color and size of nodes respectively. For larger sets of rules (i.e., all), arulesViz was used to export rules to a GraphML format for interactive visualization using tools such as Gephi\textsuperscript{37}.

As an initial validation of the results, several literature sources were reviewed for known familial risk factors and comorbidities, including chapters on asthma within pediatric textbooks\textsuperscript{39-40} and published studies or reports focused on pediatric asthma\textsuperscript{8-16,41}. Two pediatric clinical experts (RCW and PTR) also provided further interpretation of the family history associations.

### Results

Using the specified criteria, 2,048 pediatric asthma patients were identified where 882 (43.1%) were from problem list only, 68 (3.3%) from encounter diagnosis only, and 1,098 (53.6%) from both. Of these patients, 1,646 (80.4%) had at least one family history entry since 2009 and 1,531 (74.8%) had entries associated with encounters in 2014; this latter set formed the cohort for this study. Of the 10,260 most recent family history entries for this cohort (as of December 31, 2014), 7,722 (75.3%) were associated with positive family history and 2,538 (24.7%) with negative family history. After excluding entries with *“* indicated as the problem, 7,342 (71.6%) positive and 2,535 (24.7%) negative family history entries remained. Table 2 includes the distribution of entries, age associated with the most recent set of entries, and sex for the full dataset as well as the positive and negative family history subsets.

**Table 2. Distribution of Entries, Age, and Sex for Full Dataset and Positive and Negative Family History Subsets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Entries</th>
<th># Patients</th>
<th>Range</th>
<th>Mean±SD</th>
<th>Range*</th>
<th>Mean±SD</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>10,260</td>
<td>1,531</td>
<td>1–34</td>
<td>6.7±5.1</td>
<td>3.0–18.0</td>
<td>11.3±4.4</td>
<td>683 (44.6%)</td>
<td>848 (55.4%)</td>
</tr>
<tr>
<td>Positive</td>
<td>7,342</td>
<td>1,433</td>
<td>1–32</td>
<td>5.1±4.1</td>
<td>3.0–18.0</td>
<td>11.3±4.4</td>
<td>647 (45.1%)</td>
<td>786 (54.8%)</td>
</tr>
<tr>
<td>Negative</td>
<td>2,535</td>
<td>690</td>
<td>1–18</td>
<td>3.7±3.0</td>
<td>3.0–18.0</td>
<td>10.9±4.4</td>
<td>297 (43.0%)</td>
<td>393 (57.0%)</td>
</tr>
</tbody>
</table>

* maximum age = 17.999

The frequency and prevalence of family history problems for the pediatric asthma cohort were determined for the `pos_problem`, `pos_problem_relation`, and `neg_problem` datasets. Table 3 includes the top 10 family history problems, family history problems with relations, and negative family history problems ranked by prevalence.

**Table 3. Ranking of Family History Problems and Relations by Prevalence ([n] indicates ranking by frequency)**

<table>
<thead>
<tr>
<th>Family History of Problem P</th>
<th>Family History of Problem P in Relation R</th>
<th>Negative Family History of Problem P</th>
</tr>
</thead>
</table>

For each of the five datasets, pairwise association rules were generated using different thresholds for support and confidence. Figure 2 depicts the change in number of rules depending on the thresholds used for the `pos_problem` dataset. Rules generated using “low” thresholds (minimum support of 0.01 and confidence of 0.1) and “intermediate” thresholds (minimum support of 0.03 and confidence of 0.3) were selected for further review.

Figure 3 includes scatter plots of the 194 and 242 rules generated using the low thresholds for the `pos_problem` and `neg_problem` datasets respectively. The upper left quadrant highlights rules that do not occur frequently (based on support value), but have a higher \( \chi^2 \) value. The top rules based on \( \chi^2 \) for the `neg_problem` dataset involved negative family history of “Severe Sprains,” “Broken Bones,” “Collagen Disease,” “Dislocations,” and “Scoliosis” (e.g., \{Severe Sprains\} \( \Rightarrow \) \{Broken Bones\} and \{Collagen Disease\} \( \Rightarrow \) \{Dislocations\} with support=0.08 and \( \chi^2=352 \).
Figure 2: Number of rules for combinations of minimum support and confidence thresholds for pos_problem.

<table>
<thead>
<tr>
<th>support</th>
<th>support</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>18000</td>
<td>290</td>
<td>98</td>
<td>64</td>
<td>46</td>
<td>38</td>
<td>38</td>
<td>34</td>
<td>33</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>0.1</td>
<td>1691</td>
<td>194</td>
<td>85</td>
<td>61</td>
<td>45</td>
<td>38</td>
<td>38</td>
<td>34</td>
<td>33</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>0.2</td>
<td>1332</td>
<td>151</td>
<td>61</td>
<td>48</td>
<td>38</td>
<td>34</td>
<td>34</td>
<td>31</td>
<td>29</td>
<td>27</td>
<td>20</td>
</tr>
<tr>
<td>0.3</td>
<td>801</td>
<td>131</td>
<td>49</td>
<td>39</td>
<td>33</td>
<td>29</td>
<td>29</td>
<td>27</td>
<td>25</td>
<td>23</td>
<td>17</td>
</tr>
<tr>
<td>0.4</td>
<td>572</td>
<td>92</td>
<td>32</td>
<td>23</td>
<td>19</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>0.5</td>
<td>472</td>
<td>53</td>
<td>21</td>
<td>14</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>0.6</td>
<td>274</td>
<td>24</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0.7</td>
<td>215</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.8</td>
<td>197</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>187</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>187</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 3: Scatter plots showing the distribution of rules for pos_problem (a) and neg_problem (b) using minimum support of 0.01 and confidence of 0.1. Rules are plotted based on support and \( \chi^2 \) values with color reflecting confidence value (darker shading represents higher confidence value).

For the pos_problem dataset, Figure 4 includes graph-based visualizations of the 39 rules and top 50 rules generated with the intermediate and low thresholds, which demonstrate how different family history problems and rules are highlighted with varying thresholds. For example, the top rule based on \( \chi^2 \) in Figure 4a is \{[High Cholesterol] \} => \{[Hypertension] \} with support=0.17 and \( \chi^2=171.31 \) and in Figure 4b is \{[Ulcers] \} => \{[Heartburn/Reflux] \} with support=0.01 and \( \chi^2=201.64 \).

Figure 5 presents grouped matrix-based visualizations for the 9 and 31 rules generated using the intermediate thresholds for the pos_problem_relation and pos_problem_side datasets respectively while Figure 6 shows all 205 rules generated for pos_problem_degree with the low thresholds in the Gephi (0.8.2) visualization tool. These three sets of rules represent more granular associations in comparison with those from pos_problem and convey the presence of family history problems in particular relatives, side of family (maternal or paternal relative), and degree of relationship (first or second degree relative). While lower ranked, there are several rules including asthma in particular relatives or side of family such as \{[Asthma]_[Brother] \} => \{[Asthma]_[Mother] \} with support=0.03 and \( \chi^2=9.28 \) in Figure 5a and \{[Asthma]_[paternal] \} => \{[Asthma]_[maternal] \} with support=0.01 and \( \chi^2=1.34 \) in Figure 5b. In Figure 6, the highest ranking rule is \{[Arthritis]_[first] \} => \{[Arthritis]_[second] \} with support=0.01 and \( \chi^2=99.84 \) and lowest is \{[Thyroid Disease]_[first] \} => \{[Asthma]_[first] \} with support=0.02 and \( \chi^2<0.01 \).

The reviewed literature sources described several known familial risk factors for pediatric asthma: family history of atopic diseases (asthma, allergic rhinitis, and atopic dermatitis [eczema]) in general as well as in specific relatives (parents, siblings, and grandparents)\(^6\), \(^8\)-\(^10\), \(^38\)-\(^41\). Some studies also reported on the effect of number of relatives (asthma in one or more relatives) as well as sex (male or female relatives)\(^8\)-\(^10\). Asthma co-morbidities include: sinusitis, gastroesophageal reflux disease (GERD), obesity, psychological disturbances (particularly depression and
anxiety disorders), and tobacco use. In addition, conditions reported to be possibly increased with asthma include: hypertension, diabetes, heart disease, arthritis, and cancer. Collectively, the results in Table 3 and Figure 4-6 are consistent with the existing evidence of familial risk factors and co-morbidities for asthma.

**Figure 4:** Graph-based visualization of rules for pos_problem with minimum support of 0.03 and confidence of 0.3 [39 rules] (a) and minimum support of 0.01 and confidence of 0.1 [194 rules; only top 50 shown] (b). In this representation, family history problems and rules are represented as vertices and edges indicate the relationship between problems in rules. The size of the vertex represents the $\chi^2$ value (larger circle for higher value) and color corresponds with support value (darker shade for higher value).

**Figure 5:** Grouped matrix-based visualizations of rules for pos_problem_relation (a) and pos_problem_side (b) using minimum support of 0.03 and confidence of 0.3. The x-axis displays items in the LHS of the rule and y-axis displays items in the RHS of the rule. Circle size and color reflect the strength of a rule based on $\chi^2$ and support respectively.
Discussion

In this study, the feasibility of using structured family history information in the EHR was explored for identifying and assessing interactions among potential familial risk factors for pediatric asthma. The preliminary results highlight the potential of the developed approach for validating known knowledge and suggesting opportunities for further investigation. As reflected in the analyses (Figures 4-6), rules including family history of asthma, allergic rhinitis, and allergies in general, specific relatives, maternal and paternal side of family, and first- and second-degree relatives support what has been reported in the literature. Other rules reflect reported asthma co-morbidities (e.g., Heartburn/Reflux, Depression, and Anxiety Disorders) as well as common chronic conditions in adults (e.g., Cancer, Diabetes, Heart Disease, Hypertension, and High Cholesterol). There is also some evidence of associations for asthma and potentially related conditions (e.g., otitis media) with Migraines and Hearing Loss, which appeared in Figure 6. The findings from this study represent a first step towards further understanding the potential effect of family history of asthma, asthma co-morbidities, and other chronic conditions that may ultimately contribute to informing enhancements to tools for predicting increased risk of asthma in children (e.g., the Asthma Predictive Index that specifies parental asthma among the major criteria).

In performing the case study for pediatric asthma, several next steps were identified for enhancing the knowledge discovery process. Cohort identification was based on documentation of particular ICD-9-CM codes as the encounter diagnosis or on the problem list. Future studies would explore use of other data in the EHR (e.g., medications such as inhaled corticosteroids and clinical notes) and compare the accuracy of these different sources. Additional fields from the family history module in the Epic EHR at the University of Vermont Children’s Hospital as well as family history documented within clinical notes could serve to complement the structured problem and relation fields. For example, as shown in Table 1, the free-text comment field may include problems, relations, or other information that could not be documented in structured format. Existing NLP tools for family history could be adapted to extract information from the comments and notes for subsequent integration and data mining.

Several known issues are associated with association rule mining such as the generation of large numbers of rules that can present challenges for identifying meaningful associations. A related challenge is ensuring that important rules are not missed due to high thresholds (referred to as the “rare item problem”). To address these challenges, a number of algorithms and techniques have been proposed such as generalized association rule mining that involves use of concept hierarchies to generate rules at different levels of granularity. In this study, family history entries...
were analyzed at four levels of granularity where the most general rules were generated from pos_problem, most specific rules from pos_problem_relation, and intermediate rules from pos_problem_side and pos_problem_degree. Next steps include exploring use of existing concept hierarchies for relations (e.g., HL7 Vocabulary for RoleCode that includes almost 150 values organized in a six-level hierarchy) and disease groupings to also enable generalization of problems (e.g., Clinical Classifications Software and PheWAS groups for ICD-9-CM). Different thresholds for support and confidence were used to understand the effect of these constraints on number and content of rules and two sets of thresholds were selected as a demonstration (Figure 2). Further work is needed to determine the balance between number and quality of rules to facilitate interpretation and evaluation.

In interpreting the generated rules, questions arose regarding the appearance of common chronic conditions and whether there may be bias due to documentation practices. Different control populations (e.g., pediatric patients who do not have asthma or have other chronic conditions such as diabetes) could be used to compare family history associations, filter common rules, and thus highlight those that may be unique and relatively more common to the pediatric asthma population. Documentation by different providers or clinics (e.g., primary care vs. specialist) as well as variable EHR user interfaces for family history also warrants further investigation. For example, the top rules generated for neg_problem included negative family history problems that were often associated with encounters in a pediatric orthopedics clinic, speaking perhaps to differences in documentation practices and possibly how family history questions are asked. For the initial validation, a number of literature sources were used to identify known familial risk factors and co-morbidities for asthma. Biomedical literature (e.g., in PubMed/MEDLINE) and other data sources (e.g., public health surveys) could serve as potentially valuable sources for mining and discovery of associations among risk factors for diseases that could complement or be used to validate those generated from the EHR. Next steps also include exploring additional visualization techniques and tools (e.g., R packages for heatmaps) to improve readability and facilitate interpretation as well as conducting more formal evaluations that would involve categorization of rules by clinical experts (e.g., as known/unknown or direct/indirect).

The approach used in this study was designed to be generalizable to other institutions, risk factors, and conditions. Use of open-source technologies (R and Ruby) also provides a flexible, configurable, and extensible framework. Planned extensions to the approach include incorporating standards to promote knowledge sharing and comparison across institutions. For example, the local codes assigned to problems and relations used in this study could be mapped to the Unified Medical Language System (UMLS) Metathesaurus and HL7 Vocabulary for RoleCode respectively. Based on a preliminary mapping, ~93% of the problems could be mapped to UMLS concepts using a combination of MetaMap and manual searches while 73% of the relations could be mapped to HL7 Vocabulary codes. Other next steps include incorporating risk factors documented as part of the social history in the EHR to gain insights to interactions among familial, social, and behavioral factors for diseases such as pediatric asthma.

Conclusion

The widespread adoption of EHR systems has the potential to contribute to enhancing knowledge of interactions among risk factors for diseases. This study demonstrated the use of structured family history information from an EHR to identify pairwise associations representing interactions among potential familial risk factors for pediatric asthma. The preliminary findings support existing evidence and provide guidance for next steps in exploring the use of family history in disease risk assessment.

Acknowledgments

Research reported in this manuscript was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011364. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Cloud-based Predictive Modeling System and its Application to Asthma Readmission Prediction

Robert Chen, BS¹, Hang Su, BS¹, Mohammed Khalilia, PhD¹, Sizhe Lin¹, Yue Peng, BS¹, Tod Davis, BA², Daniel A. Hirsh, MD³, Elizabeth Searles, RN, BSN, MBA², Javier Tejedor-Sojo, MD², Michael Thompson, BS, MBA², Jimeng Sun, PhD¹

¹School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA
²Children’s HealthCare of Atlanta, Atlanta, GA, USA
³Pediatric Emergency Medicine Associates, LLC, Atlanta, GA, USA

Abstract

The predictive modeling process is time consuming and requires clinical researchers to handle complex electronic health record (EHR) data in restricted computational environments. To address this problem, we implemented a cloud-based predictive modeling system via a hybrid setup combining a secure private server with the Amazon Web Services (AWS) Elastic MapReduce platform.

EHR data is preprocessed on a private server and the resulting de-identified event sequences are hosted on AWS. Based on user-specified modeling configurations, an on-demand web service launches a cluster of Elastic Compute 2 (EC2) instances on AWS to perform feature selection and classification algorithms in a distributed fashion. Afterwards, the secure private server aggregates results and displays them via interactive visualization.

We tested the system on a pediatric asthma readmission task on a de-identified EHR dataset of 2,967 patients. We conduct a larger scale experiment on the CMS Linkable 2008-2010 Medicare Data Entrepreneurs’ Synthetic Public Use File dataset of 2 million patients, which achieves over 25-fold speedup compared to sequential execution.

INTRODUCTION

High healthcare costs have placed a burden on the federal budget in the United States. Over 90% of Medicare expenditure can be attributed to the management of chronic diseases.¹ However, the quality of care is still far from optimal for many patients, especially those with chronic conditions such as asthma.² While there has long been an interest in lowering asthma readmission rates, most predictive modeling studies for asthma have applied a small number of models and may be limited by small datasets. Fortunately, the rapid adoption of electronic health records (EHRs) in healthcare systems provides an exciting opportunity for researchers to leverage this data for secondary uses such as predictive modeling.

While predictive modeling approaches can aid in the detection of readmissions, the predictive modeling process is tedious and time consuming. Researchers often evaluate many models and compare performance metrics between them. Each model may involve different cohort selection criteria, or different features used in predictive modeling tasks. Furthermore, researchers may elect to evaluate several different algorithms in order to choose the best method for predicting a particular target outcome. These iterative predictive modeling efforts will accumulate and lead to large differences in performance metrics attained when comparing the outcomes of different models. Furthermore, with the tsunami of EHR data we need a more scalable computing infrastructure. Taking the aforementioned drawbacks together, we argue that the traditional predictive modeling pipeline is in need of a major overhaul.

With the rapid adoption of EHR systems in hospitals, predictive modeling will be of major interest in the clinical setting. A number of studies have performed predictive modeling for applications such as asthma readmission prediction in hospitals.³–⁶ However, most of these studies were done using either standalone software products for statistical analysis, or computer code written independently by researchers. Such approaches are often conducted entirely on the researchers’ local computers, and are not scalable with large datasets that are made available as EHR adoption grows rapidly.

Meanwhile, there is evidence that cloud computing can be leveraged in order to support big data analytics on large datasets over a large number of machines in a distributed setting.⁷,⁸ To date, there does not exist a cloud based web service that supports predictive modeling on large healthcare datasets using distributed computing. There
have been some implementations of predictive modeling software. For example, McAulley et al. built a standalone application for clinical data exploration and machine learning. However, the tool was run on local machines and was not deployed on the cloud for easy use by others. The lack of development of health analytics systems on the cloud may also partially be due to the concern of privacy and security of patient data on the cloud.

In addition to the problem with large datasets, researchers often run many iterations of predictive modeling studies before arriving at a desired result. Each iteration may involve changes in the study cohort, features used, and specific machine learning algorithms run. Constantly toggling these parts of the process is tedious and may result in errors. Ng et al. developed the PARAMO system, a predictive modeling platform which constructs a large number of pipelines in parallel with MapReduce/Hadoop. However, PARAMO is built on the user’s own cluster, which is not always available in every clinical institution, and also lacks scalability when faced with large datasets beyond the capacity of their existing cluster. In addition, most pipelines such as PARAMO are difficult to deploy in a clinical setting due to the large expenses required to maintain servers. Therefore, these systems make little to no impact on clinical decision-making.

To help address the limitations of current predictive modeling pipelines, we developed and deployed a hybrid system that combines a secure private server with the cloud-based Amazon Web Services (AWS) Elastic MapReduce platform. The system consists of a web service that runs on a private server in a secure environment for preprocess patient data into feature matrices, and an on-demand AWS web service to perform predictive modeling computations. Note that such a hybrid setup enables security of the patient data and at the same time leverages the scalable computing infrastructure on the cloud. Our system is highly customizable to support various predictive model configurations. Furthermore, the system is highly scalable, as the number of cloud-based machines launched can increase with the size of input data. Finally, the system is cost effective because the AWS Elastic MapReduce cluster is only launched when predictive modeling jobs need to be run.

We applied our platform to a predictive modeling task of identifying patients at risk for asthma readmission using a cohort of patients from the Children’s Healthcare of Atlanta EHR system. As one of the most common chronic illnesses in children, Asthma costs over $56 billion each year, placing a financial burden on the healthcare system. Asthma affects 10.5 million children in the United States annually, and leads to a total of 10.5 million missed school days each year. However, a child whose asthma is properly controlled with education, medication and lifestyle has a better chance of avoiding emergency department (ED) visits as an inpatient. In addressing these issues, care managers want to understand the trends and patterns in the entire patient population. Currently, that is done by grouping patients using diagnosis categories such as Clinical Risk Groups (CRGs) or by risk stratification using risk scores such as clinical risk scores (CRS) for asthma. However, neither approach provides homogeneous patient clusters for purposes of determining targeted treatment protocols. As an alternative, we propose to use our predictive modeling system based on a machine learning strategy to identify patients at high risk for readmission from context-specific information.

In addition to the asthma readmission prediction task, we showed that our system is scalable to large datasets by successfully running a prediction task on the publicly available CMS Linkable 2008-2010 Medicare Data Entrepreneurs’ Synthetic Public Use File (DE-SynPUF) dataset, which includes over 2 million patients.

**METHODS**

In this section we describe the design and implementation of our predictive modeling system.

**General Overview**

The system is composed of two main elements, a persistent web service in a private environment such as a hospital’s internal server and an on-demand web service on a cloud-computing environment such as AWS. The persistent web service constantly runs on a dedicated server, houses raw EHR data, and performs preprocessing of data. Processed data in the form of event sequence files are sent to the predictive modeling module, which is powered by an on-demand web service such as an AWS Elastic MapReduce cluster. Once the predictive modeling module is finished, the results are aggregated and displayed to the user in the performance analysis module running on the persistent web service. Figure 1 illustrates the key components of our system.

**Preprocessing**

Data from the EHR are first converted into event sequence files to be used as input to the predictive modeling module. The event sequence files are in the form of text files where each line represents one distinct event from the database. Each record is represented by a tuple in the format \((\text{patient, event, timestamp, value})\), where patient, event, timestamp and value represent the patient ID number, event name, date and time of the record, and a value for the event, respectively. In the case of binary events, such as medication and diagnostic events, the value for a tuple
is set equal to 1. In the case of events that are normally associated with numerical values, such as lab values, the value for the tuple is set to the numerical value that is present in the record. In the case of categorical events, such as gender recorded in an admission, the value for the tuple is set to the alphabetical value (i.e. ‘F’ and ‘M’ for gender) from the record. The final event sequence files are used as inputs to the predictive modeling module.

Instead of using the raw information such as actual patient IDs and ICD-9 codes, the data in event sequence files are further transformed into internal coded values on the persistent web service before being uploaded to the on-demand AWS web service. In particular, patient IDs and event names can be hashed into internal IDs. Thus, the raw patient information would not be used in the predictive modeling processes running on the cloud for security considerations. After the predictive modeling module is finished, specific patient ID numbers and event names may be decoded on the persistent web server before running the performance analysis module on the dedicated server.

**Predictive Modeling Module**

The predictive modeling module consists of several stages. First the cohort construction and feature construction stages are conducted. Next, cross validation stages comprised of feature selection followed by either classification or regression are run. We call a concrete step in one predictive modeling process a task. Examples of tasks include constructing diagnoses features, or building a logistic regression classifier on a specific training set. We organize all those tasks into different computation stages: cohort construction, feature construction, cross validation data splitting, feature selection, classifier training and classifier testing. Note that each stage corresponds to one or many computational tasks. All of these tasks occur on the on-demand web service, which is implemented with AWS Elastic MapReduce.

The predictive modeling module launches a new AWS Elastic MapReduce cluster consisting of multiple EC2 instances for each analysis workload. The predictive modeling module aggregates all computational tasks from all stages of predictive modeling process and schedules them to run in parallel on the provisioned AWS cluster. The user is allowed to choose the number EC2 virtual machine instances and the types of machines to use in the MapReduce cluster. Next we introduce those computation stages in more details.

**Cohort construction:** Once event sequence files are uploaded to the on-demand web service, the user can specify a set of filtering criteria that can be used for constructing cohorts (Figure 2A). A user can define the patient selection criteria for cases and controls. For example, for an asthma readmission prediction study, the user may define the case patients by inputting readmission as the target event to predict. Patients who have the target event will be regarded as cases and others as controls.

After user defines the target event to predict, the user can further narrow down the cohort to study by specifying conditions. For example, user may require all patients in cohort should have at least 3 inpatient events. After the user selects case cohort inclusion conditions, the user may want to balance the number of cases and

---

1 See http://aws.amazon.com/ec2/instance-types/ for more details
controls via matching (Figure 2A). In this situation, the user may elect to select a limited number of matched controls using a matching algorithm based upon patient similarity metrics or propensity scores. Currently, a simple case-control matching algorithm is implemented, which selects control patients that have identical or numerically similar values for the matching criteria as the case patients.

**Feature construction:** Features from the event sequence input data may be aggregated depending on user preferences and study design. The feature construction module allows the user to specify a method of aggregation for each event type with respect to its values for each patient. The aggregation methods include mean, sum, count, and latest. Mean and sum are the mean and sum of the values across all certain events for a given patient. Count is the number of times the feature occurs as an event. Latest is the value for the feature at the most recent occurrence. Figure 2B shows a screenshot of the feature construction module.

During the feature construction phase, besides the feature matrix, other entities will also be computed and stored for later usage. For example, feature value statistics will be used for the performance analysis module on the persistent web service. Feature value statistics we collect include percentage of cases/controls who has certain events and distributions of feature values within cases and controls.

**Feature selection:** The predictive modeling module runs a combination of feature selection and classification tasks. Since EHR data is often high dimensional in nature, some features may have a large amount of missing or noisy information. Feature selection is used to filter out these features, which should not be considered for predictive modeling. The feature selection algorithms implemented include raw features, the Chi-square feature selection, analysis of variance (ANOVA) F-value based feature selection, and false discovery rate based feature selection. The raw features method uses all constructed features from the data. The Chi-square method uses a Chi-squared test to evaluate p-values for each feature with respect to the target labels. A smaller p-value indicates a more predictive feature. For example, user may choose to select features with p-values less than 0.05. Likewise, the analysis of...
variance (ANOVA) F-score method applies an ANOVA test to each feature with respect to the target labels. The false discovery rate method selects features for being correlated with the target labels using the Benjamini-Hochberg algorithm.17

**Classification and regression:** We included all classification algorithms from the Python scikit-learn18 package, including logistic regression (LR), support vector machine (SVM), k nearest neighbors (KNN), random forest (RF) and beyond. The user is able to specify parameters for each classifier, for example the type of kernel for SVM, the number of trees for RF, the number of neighbors and the distance metric for KNN, and the method of regularization and type of optimization in LR. Similarly, all regression algorithms from the scikit-learn package are also included.

**Cross validation:** In the case of classification or regression problems, cross validation is run for each possible pairing of feature selection method followed by classification or regression methods. The user is able to set the number of folds in cross validation and number of iterations to run cross validation (Figure 2B). For each fold of cross validation, feature selection is first run on the training folds of the dataset. The selected features are used in the classification algorithm to be run. The system will figure out the set of tasks to run and their dependencies. For example, on a given fold of cross-validation, classifier C’s input depends on the output of feature selector S of the same fold, thus the system runs C after S. For other tasks independent from C or S, the system could schedule them running in parallel with C and S.

**Parallelization:** As described previously, our system runs on an on-demand AWS Elastic MapReduce cluster. The system achieves speed and scalability from two levels of parallelization. The first level is within-task parallelization, which means a predictive modeling task itself is implemented using big data parallel processing techniques. For example, cohort construction and feature construction are implemented using Spark19. Thus, different features will be constructed in parallel. Another level is between-task parallelization. All tasks derived from the predictive model configuration form a dependency graph, and the system schedules tasks in topological order. A task will be ready to run whenever its dependencies finish. For example, testing of a classifier could be scheduled to run when the training of the given classifier finishes. All tasks whose dependencies satisfied could run in parallel. For example, if both classifier A and B depend on feature selector C, A and B can start running simultaneously when C finishes. The degree of parallelization will be determined by the capacity of the cluster, which is usually measured by total number of CPU cores and total memory available. All the scheduling and execution are done using Hadoop and Cascading running on AWS MapReduce clusters.

Once all tasks are finished, the system collects and aggregates the results from all clusters’ distributed file systems. The results are sent to the performance analysis module on the persistent web service.

**Performance Analysis Module**

The final module of the system is the performance analysis module on the dedicated server. The results from all classifiers run in the predictive modeling module’s MapReduce service are collected, aggregated and stored in a MongoDB database running on the persistent web service. The performance metrics calculated include area under the receiver operating curve (AUC), positive predictive value (PPV), sensitivity, F1 score, accuracy and Matthews correlation coefficient. The web service retrieves the performance metric data from the MongoDB database and displays results to the user in an intuitive interactive interface. Figure 2C shows a screenshot of the webpage displayed to users.

**RESULTS**

Next we describe the results from the application of our predictive modeling platform to an asthma readmission task.

**Asthma Prediction Task Experiment Setup**

The study involved a cohort of 2,967 inpatient pediatric asthma patients from the Children’s Hospital of Atlanta (CHOA). There were 1,493 patients who had at least one readmission for asthma treatment, and 1,474 patients who did not have any readmissions. Data for inpatient events representing emergency department initial visits and readmission visits were used. Table 1 showcases general patient characteristics of the study cohort.

To run the preprocessing and performance analysis modules, we hosted the persistent web service on dedicated server running the Ubuntu 14.04.1 LTS operating system, with 24 Intel Xeon 2.6GHz processors (six cores each) and 256 GB of RAM. An on-demand web service was launched in order to run the predictive modeling module. The on-demand web service consisted of an AWS Elastic MapReduce cluster consisting of 1 master m3.medium EC2 instance and 20 slave c3.xlarge EC2 instances.

**Feature Construction:** We preprocessed the data set and obtain 5,728 unique patient visits. Each visit has a readmission label showing that whether or not the visit has led to one or more visits within the next 12 months.

We constructed six groups of features: demographic features, diagnosis features, medication features, procedure features, and visit features. The demographic, diagnosis, medication and procedure features are
categorical features while the lab and visit features are numeric features. Table 1 shows descriptions and example values for features in each group. We converted the categorical features to 1-over-K binary code representations. For example, the feature “race” has five distinct categories: White, Black or African American, Asian, American Indian or Alaska Native, and Others. If a patient belongs to the category white, his or her “race” will be represented by a 5-dimensional feature vector \([1,0,0,0,0]\). The diagnosis and medication features were binary features, where a value of 1 indicates that an event occurred for that feature and a value of 0 indicates that an event did not occur. Furthermore, we convert the numeric features to z-scores. Taking a 40-month-old patient for example, the z-score of feature “age” will be -0.91 given that the average and standard deviation of “age” in our cohort are 100.10 months and 66.13, respectively.

<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Features</th>
<th>Type</th>
<th>Example Name</th>
<th>Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td>Race</td>
<td>Categorical</td>
<td>White</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
<td>Categorical</td>
<td>Female</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>Numeric</td>
<td>40 months</td>
<td>Latest</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>ICD9 Code</td>
<td>Categorical</td>
<td>33.9</td>
<td>Count</td>
</tr>
<tr>
<td>Medication</td>
<td>Medication Name</td>
<td>Categorical</td>
<td>Albuterol</td>
<td>Count</td>
</tr>
<tr>
<td>Procedure</td>
<td>Procedure ID</td>
<td>Categorical</td>
<td>404082</td>
<td>Count</td>
</tr>
<tr>
<td>Lab</td>
<td>Lab Name</td>
<td>Numeric</td>
<td>Glucose</td>
<td>Mean</td>
</tr>
<tr>
<td>Administration</td>
<td>Length of Stay</td>
<td>Numeric</td>
<td>5 hours</td>
<td>Mean</td>
</tr>
</tbody>
</table>

Table 1: A summary of features constructed in the experiment on prediction of asthma readmission. The aggregation method used during the predictive modeling module is also reported.

**Cohort Construction:** We obtained a cohort of 1,493 unique patients with asthma readmission within one year after being discharged and 1,474 unique control patients without readmission matched on age in month and gender. We use a 1919-dimensional feature vector to represent each patient. We summarize the statistics of demographics of the cohort in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>All Patients</th>
<th>Readmission</th>
<th>No Readmission</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2967</td>
<td>1493</td>
<td>1474</td>
</tr>
<tr>
<td>Age, years (mean)</td>
<td>5.0</td>
<td>5.2</td>
<td>4.9</td>
</tr>
<tr>
<td>Gender (% male)</td>
<td>59.9%</td>
<td>59.9%</td>
<td>60.0%</td>
</tr>
<tr>
<td>Race (% white)</td>
<td>36.1%</td>
<td>36.0%</td>
<td>36.2%</td>
</tr>
<tr>
<td>Race (% black)</td>
<td>52.6%</td>
<td>52.4%</td>
<td>52.7%</td>
</tr>
</tbody>
</table>

Table 2: General patient characteristics of the study cohort. Demographic features are shown for all patients, as well as for patients with at least 1 readmission event, and patients without any readmission events.

**Feature selection:** We performed feature selection using four separate methods: raw features, ANOVA F-score feature selection, Chi-square feature selection, and false discovery rate (FDR) feature selection.

**Classification:** We formulated the asthma readmission prediction as a binary classification problem where the two target labels are defined as follows:

1: at least one readmission within 12 months of any inpatient visit
0: otherwise

We applied four commonly used classifiers: logistic regression (LR), linear support vector machine (linear SVM), K-nearest neighbor (KNN), and random forest (RF). We used stochastic gradient descent with L2 regularization for the logistic regression, set K=1 and use Euclidean distance for KNN, used a linear kernel with c=1 for SVM, and used 100 trees for RF.

**Performance analysis:** We partitioned the patients into training and testing cohorts in a 3 times 5-fold cross validation process, meaning cross validation was run for 3 iterations. For each fold, we first performed feature selection and then trained the model on the training set (80% of the entire data) using the selected features. Afterwards, we evaluated the model performance on the testing set (20% of the entire data). We used the following evaluation metrics: a) area under the receiver operating characteristic curve (AUC); b) positive predictive value.

---

ii There is a larger number of cases than controls because multiple cases can match to the same control patient.
(PPV); c) sensitivity; d) F1 score; e) accuracy. To calculate the final value for each performance metric, we find the mean of the means of each metric across all iterations.

**Asthma Experiment Results**

**Feature selection:** Of all of the feature selection methods, the false discovery rate (FDR) method achieved the best overall performance with AUC 0.69, PPV 0.69, sensitivity 0.46, F1 score 0.55, and accuracy 0.77. Table 3 shows the top predictive features selected by the FDR feature selection method in all 10 folds. Six out of the 10 features were verified by pediatric clinicians to be possible indicators for asthma readmission (highlighted in Table 3). Two of the features, the medication *fluticasone-salmeterol* and the lab *total immunoglobulin E (IgE)*, are known to be strong indicators for asthma readmission. The *fluticasone-salmeterol* feature is present in 15% of all cases while present in only 8% of all controls. This result is clinically meaningful because *fluticasone-salmeterol* is commonly prescribed in more severe asthmatic patients. The *total immunoglobulin E (IgE)* lab value is 565 IU/mL in cases and 258 IU/mL in controls. This result is clinically meaningful as well, since more severe asthmatic patients tend to have higher values for IgE, a marker indicating sensitivity to allergens.23

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Percent (case)</th>
<th>Percent (control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medication</td>
<td>Montelukast</td>
<td>30.3</td>
<td>19</td>
</tr>
<tr>
<td>Medication</td>
<td>Fluzone</td>
<td>2.6</td>
<td>1</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Extrinsic asthma with status asthmaticus</td>
<td>26</td>
<td>14.1</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Other pulmonary insufficiency, not elsewhere classified</td>
<td>8.7</td>
<td>3.8</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Contact dermatitis and other eczema, unspecified cause</td>
<td>22</td>
<td>13.1</td>
</tr>
<tr>
<td>Medication</td>
<td><em>Fluticasone-salmeterol</em></td>
<td>15.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Medication</td>
<td>0.9% sodium chloride (PF)</td>
<td>14.4</td>
<td>8.4</td>
</tr>
<tr>
<td>Medication</td>
<td>D5-1/2NS</td>
<td>23.2</td>
<td>16.7</td>
</tr>
<tr>
<td>Lab</td>
<td>Point of care hemoglobin test</td>
<td>6.9</td>
<td>4.7</td>
</tr>
<tr>
<td>Lab</td>
<td><em>Total Immunoglobulin E (IgE)</em></td>
<td>1.6</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 3: The top 10 most predictive features selected by univariate feature selection based upon ANOVA F-value. Features verified by clinicians to be possible indicators for asthma readmission are shown in bold print.

**Classification:** We performed cross-validation to choose the appropriate number of features that gives the best performance. Cross validation was performed on each possible combination of feature selection algorithm and classification algorithm. For each feature selection method, we collected all features that met the feature selection criteria. These features were used as predictive features in the classification tasks. Table 4 shows the performance of the linear SVM classifier while using different feature selection methods and raw features. The feature selection method with the highest average of all performance metrics was determined to be the one with the best performance.

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>0.63 (0.01)</td>
<td>0.70 (0.01)</td>
<td>0.33 (0.01)</td>
<td>0.45 (0.01)</td>
<td>0.75 (0.004)</td>
</tr>
<tr>
<td>Chi-square</td>
<td>0.63 (0.01)</td>
<td>0.69 (0.03)</td>
<td>0.32 (0.01)</td>
<td>0.44 (0.01)</td>
<td>0.75 (0.01)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.69 (0.01)</td>
<td>0.69 (0.02)</td>
<td>0.46 (0.01)</td>
<td>0.55 (0.01)</td>
<td>0.77 (0.01)</td>
</tr>
<tr>
<td>All features</td>
<td>0.69 (0.01)</td>
<td>0.69 (0.02)</td>
<td>0.46 (0.01)</td>
<td>0.55 (0.01)</td>
<td>0.77 (0.01)</td>
</tr>
</tbody>
</table>

Table 4: Performance of linear SVM with different feature selection algorithms. Feature selection algorithms used include: ANOVA F-value, Chi-square, false discovery rate, false positive rate, and all features. Values shown are mean (standard deviation) across all iterations and folds of cross validation.

The results of the four different classifiers using the feature selected by the FDR feature selection method are shown in table 5. There was variability in performance of the classifiers. Linear SVM achieved the highest AUC (0.69). Logistic regression achieved the highest sensitivity (0.99), while random forest achieved the highest PPV (0.89), F1 score (0.74), and accuracy (0.86).

It is important to consider these results in the context of the particular application. For the asthma readmission prediction problem, the SVM, logistic regression, or random forest methods may all be considered effective models based upon different use cases. In cases where sensitivity may be important (e.g., detecting high risk patients who may need urgent care), logistic regression may be the best model. In cases where positive predictive value may be
important (e.g., when treatment for positively predicted patients is expensive, and financial resource allocation is important), then random forest may be the best model.

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>Linear SVM</th>
<th>KNN</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.55 (0.004)</td>
<td>0.69 (0.01)</td>
<td>0.52 (0.01)</td>
<td>0.56 (0.01)</td>
</tr>
<tr>
<td>PPV</td>
<td>0.31 (0.002)</td>
<td>0.69 (0.02)</td>
<td>0.54 (0.03)</td>
<td>0.89 (0.01)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.99 (0.001)</td>
<td>0.46 (0.01)</td>
<td>0.40 (0.01)</td>
<td>0.63 (0.02)</td>
</tr>
<tr>
<td>F1</td>
<td>0.47 (0.002)</td>
<td>0.55 (0.01)</td>
<td>0.46 (0.02)</td>
<td>0.74 (0.01)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.32 (0.004)</td>
<td>0.77 (0.01)</td>
<td>0.71 (0.01)</td>
<td>0.86 (0.002)</td>
</tr>
</tbody>
</table>

Table 5: Performance metrics for four classification algorithms implemented on features selected using the false discovery rate univariate feature selection method. Metrics reported include area under the receiver operating characteristic (AUC), positive predictive value (PPV), sensitivity, F1 score, accuracy and Matthews correlation coefficient. Values shown are the mean (standard deviation) across all iterations and folds of cross validation.

System Scalability
To demonstrate the scalability of our system, we ran our system on a much larger dataset, a set of 2.1 million patients from the CMS Linkable 2008-2010 Medicare Data Entrepreneurs’ Synthetic Public Use File (DE-SynPUF), a publicly available synthetic dataset. It contains approximately 300,000 different kinds of events from the patients. Thus, there were about 2.1 million patients and 300,000 features in the input dataset. The raw event sequence input file size is 6.88GB. We created a predictive modeling workload with more than 3000 tasks. The feature selection and classifier settings are almost identical to those used in the asthma readmission prediction task. The on-demand web service is composed of 10 r3.2xlarge AWS EC2 virtual machines. Figure 3 shows the timeline of parts of the task run and the amounts of time spent. the CMS data. The entire running time of the pipeline workflow is about 3 hours. To serve as a baseline, we ran all the tasks sequentially on a single server of the same machine configuration to calculate the total running time of a sequential run. We find that our system achieves a 36-fold speedup over the baseline sequential running time. Note that the feature construction step is only conducted once, while the data splitting, feature selection, model training and model testing steps are done for each iteration of cross validation.

**Figure 3:** Timeline of modules run and elapsed time. The data-splitting, training and testing times refer to the run times for each respective step of cross validation. Times are shown in seconds (s).

**DISCUSSION**
We have developed a cloud based system for clinical predictive modeling. Our system is the first of its kind to date, and leverages Amazon Web Services’ Elastic MapReduce technology to run distributed feature selection and classification jobs in a time efficient manner.

**Challenges in Privacy and Security**
While the usage of the cloud is relatively new, many users are already using the cloud for hosting personal health information (PHI).\(^1,7,24-26\) In the case that the users are unwilling to store EHR information into the cloud, our system architecture can be used in such a way that preprocessing of data can be performed on the persistent web server, and PHI can be mapped to codes. For example, all information including patient ID numbers and medication, procedure, lab and diagnosis names may be hashed and mapped to different values such that the raw data were not uploaded to the cloud based system.

Our system mitigates potential concerns regarding privacy and security of healthcare data. However, we also recognize the heuristic nature of our approach, and in the future we plan to conduct more focused studies on privacy consideration using the cloud in an effort to provide a more theoretical guarantee of privacy.

CONCLUSION

We have proposed and implemented a hybrid version of predictive modeling system, which combines a private dedicated instance and public cloud computing services. In this system, raw EHR data are converted into standardized features written into event sequence data files via persistent web services on the private server. The de-identified event sequence files are uploaded to an on-demand web service via Amazon Web Services, which subsequently constructs cohorts and features and schedules a series of distributed predictive modeling tasks using big data systems such as Spark and Hadoop. The results of the predictive modeling tasks are collected and displayed to the user in a highly intuitive, interactive user interface on the private server.

We applied our system to a specific task of prediction for pediatric asthma readmission using a cohort of case patients with asthma readmission and matching control patients. The predictive modeling module was successful in the prediction task through a 5-fold cross validation scheme. The system predicted patients at risk for 12-month asthma readmission with an AUC of 0.69.

We plan to improve upon the system by expanding the suite of cohort construction strategies, feature selection algorithms and classification algorithms. Furthermore, we plan to add functionality for testing multi-class classification tasks (e.g., to be used for detecting multiple types of readmissions).

ACKNOWLEDGMENT

This work was supported by the National Science Foundation, award #1418511, Children's Healthcare of Atlanta, CDC i-SMILE project, Google Faculty Award, AWS Research Award, Microsoft Azure Research Award, and the National Institutes of Health award T32-GM105490, a Computational Biology and Predictive Health Genomics Training Program at the Georgia Institute of Technology.


Inferring Clinical Workflow Efficiency via Electronic Medical Record Utilization

You Chen, PhD¹, Wei Xie, MS¹, Carl A Gunter, PhD², David Liebovitz, MD³, Sanjay Mehrotra, PhD³, He Zhang, PhD⁴, and Bradley Malin, PhD¹
¹Vanderbilt University, Nashville, TN; ²University of Illinois, Urbana, IL; ³Northwestern University, Chicago, IL; ⁴University of South Florida, Tampa, FL

Abstract

Complexity in clinical workflows can lead to inefficiency in making diagnoses, ineffectiveness of treatment plans and uninformed management of healthcare organizations (HCOs). Traditional strategies to manage workflow complexity are based on measuring the gaps between workflows defined by HCO administrators and the actual processes followed by staff in the clinic. However, existing methods tend to neglect the influences of EMR systems on the utilization of workflows, which could be leveraged to optimize workflows facilitated through the EMR. In this paper, we introduce a framework to infer clinical workflows through the utilization of an EMR and show how such workflows roughly partition into four types according to their efficiency. Our framework infers workflows at several levels of granularity through data mining technologies. We study four months of EMR event logs from a large medical center, including 16,569 inpatient stays, and illustrate that over approximately 95% of workflows are efficient and that 80% of patients are on such workflows. At the same time, we show that the remaining 5% of workflows may be inefficient due to a variety of factors, such as complex patients.

Introduction

The success of an electronic medical record (EMR) system implementation and its subsequent adoption by users is contingent upon the initial design and refinement of workflows in the healthcare organization (HCO)¹,². The appropriate design and management of workflows can significantly improve efficiency of clinical treatment³,⁴, care quality⁴, patient safety⁵, and care decisions⁶. Yet, despite their potential, workflows can be quite complex⁷, creating barriers to EMR system utilization⁸. This can ultimately lead to inefficiency in diagnoses, ineffectiveness of treatment plans, and uninformed management of an HCO. To mitigate workflow complexity in EMR systems, it has been suggested that HCOs design workflows to optimize business processes or manage the complexity of current workflows, rather than rely upon current workflows¹,⁷,⁸,⁹.

To enable such strategies, various approaches have been developed to measure the gap between the workflows defined by an HCO and the actual processes followed by individuals in a clinical setting. In general, for the gap analysis approaches, the HCO workflows are compared to the expectation of experts in the organization, as learned through surveys, interviews or observational data collected in the physical healthcare setting⁷,⁸,¹²,¹⁴,¹⁵,¹⁶. Application of these methods requires a substantial exertion of manual effort because it requires invasive interviews and patience while observing the interactions between care providers and patients. As a result, this type of approach is often limited to specific areas of clinical care. Furthermore, these methods only measure the gap between the organization’s workflows and the expectation of care providers. This neglects the influences of EMR systems on the utilization of workflows, which could be leveraged to optimize workflows facilitated through the EMR systems.

By contrast, most recent approaches model workflows by mining data recorded in the EMR system. This type of strategy considers the influences of EMR systems on the utilization of workflows, but it only models the patterns of care paths, and neglects the efficiency management of workflows⁹-¹¹.

Good management of workflow efficiency can improve quality of clinical care and reduce costs of patients¹,¹. By providing HCOs with such knowledge of workflows, we anticipate healthcare administrators will be able to optimize the efficiency, as well as minimize the complexity of workflows in a more productive manner. For instance, imagine that an HCO administrator learns a particular clinical process tends to require a long duration in time, but that the process has a large variance in its duration across the enterprise. Based on this knowledge, they can investigate the reasons behind such long waiting times and variation. In this paper, we introduce a framework to model clinical workflows at two levels of granularity and categorize these workflows according to their efficiency. To the best of our knowledge, this is the first approach to automatically learn and categorize workflows according to their duration.

416
Background
The past several years have witnessed a number of investigations into modeling and characterizing the workflows associated with clinical practices. As alluded to earlier, we characterize these investigations, and the approaches upon which they are based, into two classes. These classes are dependent upon the methods of data acquisition and analysis that are invoked: i) physical observation studies (in the clinic) and ii) virtual observation studies (data in the EMR). Here, we take a moment to review representative research from both categories and illustrate the relationship with our own approach.

Workflows Based on Physical Observations
Observation-driven studies often rely on manual data collection approaches, such as observations and interviews. One such example was presented by Unertl and colleagues, which analyzed direct observations and interviews in hospitals to understand workflow and information flow in the care of chronic diseases. Similarly, Ramaiah and colleagues designed surveys consisting of questions, interviews of care providers and patients to discover workflows associated with time delays in the HCO. In another setting, comparative data were collected from the operating room and statistical analysis was performed with respect to gains in efficiency. The ultimate goal was to justify the need for reorganizing clinical workflow to increase throughput in the operating room. In other work, ethnographic observation and interview data were applied to study the evolution and management of medical errors.

Workflows Based on Virtual Observations
Modern studies are increasingly turning to EMR-related data because they can enable large scale analytics at a low acquisition cost. While more comprehensive surveys exist about workflow mining in the EMR, we briefly examine investigations relevant to our approach.

In one study, Zhang and colleagues utilized EMR usage logs to model patients’ flow in the healthcare system. After learning patterns of patient record usage, deviations from the average workflows were detected and promoted for investigation as either undocumented policies or misuse of the EMR system. Similarly, another study employed sequence alignment methods to derive a consensus workflow and automatically detect outliers from surgical activity logs. Other studies have focused on contrasting treatment differences for certain diseases. For instance, Mans and colleagues studied stroke care by applying process mining to clinical data. They additionally compared the pathways from disparate healthcare systems and various types of patients. Most recently, Partington and colleagues focused on cross-hospital process mining and performed a comparative analysis by leveraging a combination of administrative and clinical data. This investigation yielded detailed insights into patient care and hospital budget pressures.

Methods
We provide a framework to 1) learn workflows and 2) categorize the workflows into four general types according to the length and variation in their temporal duration. For reference purposes, we summarize the common terms and notation used in this paper in Table 1.

**Event.** An event corresponds to the smallest granularity associated with a workflow. This corresponds to an action invoked by an EMR user over a patient’s medical record at a certain time. For instance, a user, acting as a pathologist, can initiate an event by accessing a patient record to request a lab test. Alternatively, another user, acting as a primary care provider, can initiate an event by accessing a patient record to approve a request to refill a medication.

**Sequence.** A patient sequence consists of a series of ordered events that represents an episode of a patient process. For example, an ordered series of events could be: a user, acting as a physician requested a lab test for a patient → a laboratory user uploaded a lab test result for the patient → the lab test results were returned to a care provider in a physician office → a registered nurse provided customer service support to the patient in response to an inquiry about the lab test results. We assume we are provided with m events that are classified into n patient sequences.

**Block.** A block is an ordered series of events that have strongly ordered relations with each other. For example, if the relation between an event Rehab Service Clinician and another event Rehab Quality Audit is strong in a way we will define precisely below, then Rehab Service Clinician → Rehab Quality Audit belongs to a block. A block represents a specific stage of a patient process, and a patient sequence consisting of events can be represented using the corresponding blocks. For instance, imagine there are two blocks [Primary Care Physician → Laboratory Tester] and [Physician Office Care Provider → Primary Care Staff Nurse]. Then, the earlier patient sequence example (with four
events) can be represented using these blocks to express a specific stage of a workflow corresponding to the four events.

**Topic.** A topic consists of a set of blocks, which together represent the main processes associated with a patient type. For example, imagine there is a group of patients with the conditions Urinary Tract Infection, Other Specified Retention of Urine, and Unspecified Essential Hypertension. Then a topic consisting of the following blocks may characterize this group: [Advanced Practice Clinician – CPOE → Physician Office], [SN-OR RN SC → Patient Care Assistive Staff], [Physician Office → Advanced Practice Clinician – CPOE], [OR RN SC-Primary → Primary Assistive Staff], [Unit Secretary → Rehab PT, Respiratory → SN-RN/Customer Service], and [SN-RN/Customer Service → NMH Physician Hospitalist-CPOE].

### Table 1. Common notation used in this study with their corresponding definitions.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E = {e_1, e_2, ..., e_n}</td>
<td>A set of events</td>
</tr>
<tr>
<td>S = {s_1, s_2, ..., s_n}</td>
<td>A set of sequences</td>
</tr>
<tr>
<td>s_i = [e_i, e_j, e_k, ..., e_n]</td>
<td>A sequence, which consists of a series of events in order.</td>
</tr>
<tr>
<td>B = [b_1, b_2, ..., b_n]</td>
<td>A set of blocks</td>
</tr>
<tr>
<td>b_i = [e_i, e_j, e_k, ..., e_n]</td>
<td>A block, which is a series of events in order.</td>
</tr>
<tr>
<td>s'_i = [b_1, b_2, ..., b_n]</td>
<td>A revised sequence, which consists of blocks.</td>
</tr>
<tr>
<td>s' = {s'_1, s'_2, ..., s'_n}</td>
<td>A set of revised sequences</td>
</tr>
<tr>
<td>T = {t_1, t_2, ..., t_n}</td>
<td>A set of topics</td>
</tr>
<tr>
<td>R_E</td>
<td>An asymmetric matrix representing the relations between events</td>
</tr>
<tr>
<td>R_B</td>
<td>A symmetric matrix representing the relations between blocks</td>
</tr>
<tr>
<td>R_E x B</td>
<td>A matrix representing the relation between sequences in S and blocks in B</td>
</tr>
</tbody>
</table>

We learn workflows at a fine granularity, in the form of blocks of events, and a coarse granularity, in terms of topics of blocks. Representation at the block-level characterizes the stage of a patient process. By contrast, representation at the coarse-grain characterizes different types of patient processes. Fig. 1 shows an example consisting of 4 models to learn workflows at the block- and topic-level. These are formalized in the Workflow Mining Algorithm (WMA) depicted in Fig. 2.

#### Figure 1. Generation of fine-gained (blocks) and coarse-gained (topics) workflows from patient sequences.

1. **Event Relation Model (ERM):** Measures relations between events according to the patient sequences. The details of relation measurement are depicted in steps 1 through 5 of WMA.

2. **Block Generation Model (BGM):** Generates blocks of events according to the relations of events and patient sequences. The details of block generation are described in steps 6 through 19 of WMA.

3. **Block Relation Model (BRM):** Measures the relations between blocks according to the common events they contain. The details about measurement are in steps 20 through 22 of WMA.

4. **Topic Generation Model (TGM):** Generates topics of blocks to represent a similar type of patient processes. The details are depicted in steps 23 to 26 of WMA.
We introduce these models in detail in the following sections on workflow generation at the block- and topic-level.

**Block-level Workflows**

First, we generate event blocks (i.e., workflows at the block-level) according to steps 1 through 19 of WMA. A workflow at the block-level aims to characterize the efficiency of a stage of a patient process. In this work, we aim to learn four types of blocks:

1. **Stable Efficient Blocks (SEB):** Have short average duration with small variance across the patient population.
2. **Unstable Efficient Blocks (UEB):** Have short average duration with large variance across the patient population.
3. **Stable Inefficient Blocks (SIB):** Have long average duration with small variance across the patient population.
4. **Unstable Inefficient Blocks (UIB):** Have long average duration with large variance across the patient population.

We anticipate that the categorization of blocks into these types can assist HCOs in speeding up the discovery of (in)efficient blocks and refine their policies accordingly. For instance, imagine the block [Radiology Mgr/RC → Attending Physician/Provider] exhibits a large variance in its duration, such that it requires less than 1 hour for one patient, but 240 hours for another patient. Our model could promote this block to administrators for investigation.

We infer blocks from the event logs generated by EMR systems. This is because EMR-facilitated workflows, and the utilization logs in particular, do not necessarily follow the exact order of events in the physical world. For instance, in the real world, the order of events may be $e_1 \rightarrow e_2$, but in some cases, the order recorded by an EMR may be $e_1 \rightarrow e_k \rightarrow e_2$. To relax the order relations of events, we consider relations within a sliding time window. Specifically, we assume that if an event and its following events are within a window of size $\alpha$, the order relations holds, but the strength of the relation is proportional to their distance. We measure the order relation between a pair of events within a sequence as:

$$\text{Event}_{\text{relation}}(e_i,e_j) = \begin{cases} \frac{1}{D_{(p(e_j) - p(e_i))^2}} & (0 < p(e_j) - p(e_i) \leq \alpha) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where $p(e_i)$ is the position of the event $e_i$ in a sequence. The position of the first event of a sequence is set as 1, and the position of the last event of a sequence is set as the length of a sequence. The relations of events are incrementally measured over all of the sequences (as depicted in steps 1 through 5 of WMA). As a consequence, this type of relation considers the ordinal distance between the events and their frequency in the patient population. When we set $\alpha$ to 1, the relation between $e_1 \rightarrow e_2$ is 4 $(1+1+1+1)$ accumulated from sequence $s_1, s_2, s_3,$ and $s_4$ respectively (as shown in Fig.1(b)).

To represent each stage of a patient process via temporal patterns, we infer event blocks through relations of events. The generation of such blocks is shown in steps 7 through 19 of WMA. We assume that, for a sequence of events, if an event has a strong relation with the following events, then they should be grouped into a block. Since the relations of events are already considered in the sliding time window, we only consider the immediately following neighbor in block generation. As shown in steps 10 through 13 of WMA, if the relation of every neighboring event is within a range of $[\beta_1, \beta_2]$, these events are included in a block. The lower bound for the range is applied to filter out weak relations, whereas the upper bound filters highly frequent relations (like stop words, “a”, “of”, or “the” in natural language processing). The end of a block is realized when the relation between an event and its neighbor is outside of the range. For instance, when $\beta_1 = 2$ and $\beta_2 = 4$, sequence $s_1 \rightarrow e_1 \rightarrow e_2 \rightarrow e_3 \rightarrow e_2 \rightarrow e_4$ (shown in Fig.1(a)) generates two blocks $e_1 \rightarrow e_2$ and $e_3 \rightarrow e_4$ (shown in Fig.1(c). This is because the event relations of $e_2 \rightarrow e_2$ and $e_2 \rightarrow e_3$ occur only once, which is below the lower bound threshold of 2.

To evaluate and characterize the time efficiency of blocks, we need contextual information, such as the duration of a block or the reasons for each event in a block. Note, a block can appear multiple times in different patient sequences, such as when $b_{1} \rightarrow e_1 \rightarrow e_2$ exists in $s_1, s_2, s_3,$ and $s_4$. As such, an event block can have multiple reasons and time durations. For example, if $e_1$ happens at 9:00am, and the following event $e_2$ happens at 10:00am in sequences $s_1$, then the time duration for $b_1$ in $s_1$ is 1 hour.

We summarize the efficiency of a block using several basic statistics. Specially, we compute the duration and variance for a block as the average and standard deviation over all occurrences in the patient population. For example, if the time duration of block $b_1$ in the four sequences is 1, 2, 1, and 1.5 hours, respectively, then the average time duration is 1.375 hours and the variance (standard deviation) is 0.4787 hours.
Figure 2. Pseudocode for the algorithm to generate workflows at the block- and topic-level.

**Topic-level Workflows**

A workflow at the block-level only provides a description for a particular stage of a patient process. However, it neglects the relations between blocks within a patient process. To do so, we summarize collections of blocks using a topic modeling strategy (e.g., latent Dirichlet allocation (LDA))\(^{18,19}\), a popular approach to learn latent concepts from a corpus of documents. In our setting, this corresponds to learning a set of latent workflow patterns to represent patient sequences. Conceptually, patient sequences can be thought of as documents, where the event blocks constitute a vocabulary and the specific event blocks assigned to a patient’s sequence are the semantic ideas derived from the vocabulary.

To learn topics, we begin by generating the set of unique blocks. We then represent each original patient sequence \(s_i\) as a new sequence \(s'_i\), consisting of blocks instead of events (step 15 of WMA). For instance, Fig.1(d) depicts 6 unique blocks generated from Fig.1(c). Now, to infer topics of terms (i.e., blocks), we need to prepare the documents (i.e., patient sequences). Here, a term is a block and a document is a patient sequence \(s'_i\) that consists of a series of blocks. While the events within a block are ordered, the order is not necessarily consistent with the real world. As such, we group the blocks associated with similar sets of events. In many respects, this can be thought of as creating a set of synonyms in a vocabulary.

To discover synonyms of blocks, we measure the similarity using the Jaccard coefficient between the blocks (as invoked in step 21 of WMA):

\[
Block_{relation}(b_i, b_j) = \frac{\text{set}(b_i) \cap \text{set}(b_j)}{\text{set}(b_i) \cup \text{set}(b_j)}
\]

where \(\text{set}(b_i)\) corresponds to the set of unique events in a block. Note that this function ranges from 0 (no relation) to 1 (perfect relation). For instance, the relation between block \(b_1\) and \(b_2\) (as shown in Fig.1(d)) is 1 (as shown in Fig.1(e)).
As described earlier, we want each patient sequence, including the synonyms of its blocks, to improve the quality of learned topics through the LDA model. We use Equation 3 to represent a patient sequence using its blocks, along with their synonyms, as:

\[
SB_{relation}(s_i', b_j) = \begin{cases} 
  1 & b_j \in s_i' \\
  1 & b_j \notin s_i' (\exists b_k \in s_i', R_{relation}(b_k, b_j) = 1) \\
  0 & \text{otherwise}
\end{cases}
\]  

(3)

We generate a matrix \( R_{s_i' \times B} \) to represent the relation between a sequence \( s_i' \) and a block \( b_j \) (steps 23 through 25 of WMA). For instance, \( s_i' \) has block \( b_1 \) and \( b_4 \) (as shown in Fig.1(c)), but \( b_1 \) has synonyms \( b_2 \) and \( b_3 \) according to Equation 3. As such, \( s_i' \) is represented by blocks \( b_1, b_2, b_3, \) and \( b_4 \) (as shown in Fig.1(f)). The matrix \( R_{s_i' \times B} \) serves as the input to the LDA learning process (step 26 of WMA).

A topic consists of a probability distribution over a set of blocks as shown in Fig.1(g). The larger the probability of a block, the more this block is representative of the topic. At the same time, the patient process can be characterized by inferred topics as shown in Fig.1(h).

Results

We evaluate our framework on four months of inpatient event logs generated by the Northwestern Memorial Hospital (NMH). In this dataset, an event corresponds to a chart access, each of which is associated with the user and the user-designated reason for the access. It should be noted that the initial reasons selected by a chart user during the hospitalization of a patient persists throughout the hospitalization. There are 1,138,317 total events distributed over 16,569 patient processes. These events were generated by 144 user roles with access to 142 reasons. Additionally, each patient is associated with a set of ICD-9 codes assigned after discharge. The total number of unique ICD-9 codes for this set of patients is 4,543.

To apply WMA, we need to set three parameters: i) the sliding window size \( \alpha \), ii) the event relation thresholds \( \beta_1 \) and \( \beta_2 \), and iii) the number of topics for the LDA model. We set \( \alpha \) to the average number of events that transpired during a 24 hour window. This corresponded to 3 events. \( \beta_1 \) was set to 50 because smaller values led to an extremely large number of blocks (over 100,000), which suggested a substantial amount of noise. \( \beta_2 \) was set to 500 because, at this point, the only events in the resulting blocks corresponded to either: 1) Physician-CPOE, 2) Residence, 3) Patient Assistive Staff, 4) Patient Care Staff Nurse, 5) Respiratory, and 6) Unit Secretary. These extremely frequent blocks limited the generation of meaningful blocks and, thus, we removed these from further consideration. For the LDA model, we set the number of topics according to topic similarity instead of perplexity, based on a previous study\(^{19}\). In doing so, we searched for a model that minimizes the workflow topic similarity, measured as the cosine similarity. We set the number of topics as 15, 20, 25, 30, and 35 respectively, and calculate the corresponding topic similarity in each setting. The topic similarity was minimized (0.0033) when the number of topics was set to 25. Given these parameters, WMA generated 22,442 event blocks and 25 topics of blocks.

In the remainder of this section, we first show the block- and topic-level efficiency results. To illustrate the results at each level, we then provide a case study with respect to the learned workflows.

Block Types

For each block, we record the average and variance in time duration. Figure 3 depicts the 22,442 blocks as a function of these concepts. We partition this space, based on their length and stability, into the four types mentioned above. To do so, we partitioned the system using thresholds of an average length of 100 hours and a variance of 100 hours. According to this split, 94.7% of the blocks are in the SEB area (i.e., short duration with low variance). This suggests that the HCO mainly manages inpatients associated with short processes.

At the same time, this finding implies that approximately 5% of the blocks are of potential concern. Among the remaining blocks, 70% correspond to SIB (i.e., long duration with low variance), 22% correspond to UIB (i.e., long duration and long variance), and 8% correspond to UEB (i.e., short duration and large variance). It is possible, however, that these blocks are artifacts of an insufficient amount of evidence to draw meaningful conclusions. To assess this issue, Figure 4 provides a frequency analysis for all of the blocks as a function of their average duration and variance. Clearly, the frequency of the efficient blocks (i.e., SEB, UEB) is substantially larger than the inefficient blocks (i.e., SIB, UIB). In most cases, the frequency of SIB and UIB blocks is small (around 2~3), which indicates they are not popular in the management of inpatients.
Block Type Case Studies

To gain a deeper appreciation of the different block types, we provide examples of UEB and UIB in Table 3. Block B1 belongs to UEB and has 6 unique reasons associated with it. Its reason R1 is associated with two significantly different durations. The first is less than 1 hour, while the other is 240 hours. The same phenomenon occurs for the reason R4, which has a duration of less than 1 hour and another of 160 hours. These phenomena illustrate how a block with the same chart access reason, can exhibit significantly different durations. They indicate that, though a block encompasses the same transitions between reasons, the time allotted for doing so may be significantly different. This may stem from a number of complications, such as varying patient symptoms and purposes for the hospitalization, the urgency of imaging needs, the ability of some attending physicians to rely on residents in training to access charts for them and provide relevant updates, the resource allocation strategies of HCOs, or the workflow timing of chart access by care providers.

Table 3. Examples of UEB and UIB block types.

<table>
<thead>
<tr>
<th>Block</th>
<th>Duration (hours)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Variance</td>
</tr>
<tr>
<td><strong>B1</strong></td>
<td>[RAD - Mgr/RC → NMH Physician-CPOE]</td>
<td>25.6</td>
</tr>
<tr>
<td><strong>B2</strong></td>
<td>[NMPG MD → CPOE → NMPG APRN → NMPG MD – CPOE]</td>
<td>203</td>
</tr>
</tbody>
</table>

Table 4 shows two patients associated with Block B1. The condition for one patient is related with polyneuritis, while the condition for the other patient is related with septicemia, which is more complex. This may be the reason why this block exhibits high variance in time duration, even for the same chart access reason.

Block B2 belongs to UIB, and its reason R2 has long duration and large variance. As can be seen, the patients associated with this block are related to obstetric care. Furthermore, advanced practice nurses function as care providers, which includes the ability to create and sign chart orders, lessening the need for rapid access by a supervising attending.

Topic Workflow Types

To characterize the categories of workflows at the topic-level, we inferred 25 topics through distribution of blocks on patients, and then categorize the 25 topics of blocks into four groups using thresholds of a duration of 20 hours and a variance of 20 hours, as shown in Figure 5. Each topic is composed of top 10 blocks with highest probabilities. The duration and variance of a topic were calculated based on average durations of top 10 blocks. Topics 1, 7, 19, 20 were classified as Unstable Inefficient Topics (UITs), topic 8 was a Stable Inefficient Topic (SIT), topic 4 and 11 were Unstable Efficient Topics (UETs) and the remaining 18 topics were Stable Efficient Topics (SETs).

In Figure 5, it can be seen that topic 20 exhibits the longest time duration and corresponding variance. This is because one of the top 10 blocks of topic 20 belongs to UIB. This block corresponds to [NMH Physician Office – CPOE → SN-RN/Customer Service → NMH Physician Office – CPOE → SN-RN/Customer Service], which has 8 different reasons. The duration and variance of this block is 300 and 170 hours, respectively.
Table 4. Examples of patients associated with B1 and B2.

<table>
<thead>
<tr>
<th>Block</th>
<th>ICD-9 Codes</th>
<th>Description of ICD-9 Codes</th>
<th>Reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>Patient 1: 340, 35781</td>
<td>Multiple sclerosis; Chronic inflammatory demyelinating polyneuritis</td>
<td>R1: Radiology Mgr/RC (a non-physician manager within the radiology department) → Attending Phys/Prov (the primary physician responsible for an inpatient’s care) [appeared twice: one time for less than 1 hour, and one time for 240 hours]</td>
</tr>
<tr>
<td></td>
<td>Patient 2: 78552, 7907, 99592, 0389, 0417, 2760, 2762, 2875, 5070, 51881, 5849, 6826, 68601, 70705, 70719</td>
<td>Septic shock; Bacteremia; Severe sepsis; Unspecified septicemia; Pseudomonas infection in conditions classified elsewhere and of unspecified site; Hyperosmolality and/or hypernatremia; Acidosis, Thrombocytopenia, unspecified; Pneumonitis due to inhalation of food or vomitus; Acute respiratory failure; Acute kidney failure, unspecified; Cellulitis and abscess of leg, except foot; Pyoderma gangrenosum; Pressure ulcer, buttock; Ulcer of other part of lower limb</td>
<td>R2: Radiology Mgr/RC → Resident-Inpatient Consulting Service [low duration, low variance]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R3: Radiology Mgr/RC → Approved Quality or Peer Review Process [low duration, low variance]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R4: Radiology Mgr/RC → Patient Care (associated with nursing roles) [appeared twice: one time for less than 1 hour, and one time for 160 hours]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R5: Radiology Mgr/RC → Resident-Inpatient Covering Service [low duration, low variance]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R6: Radiology Mgr/RC → Resident-Inpatient Primary Service [low duration, low variance]</td>
</tr>
<tr>
<td>B2</td>
<td>Patient 3: 65971, 66411, V270</td>
<td>Abnormality in fetal heart rate or rhythm, delivered, with or without mention of antepartum condition; Second-degree perineal laceration, delivered, with or without mention of antepartum condition</td>
<td>R1: Other Phys/Prov → NMSP APRN → Other Phys/Prov [long duration, low variance]</td>
</tr>
<tr>
<td></td>
<td>Patient 4: V270, 64891, 66541, V0251</td>
<td>Outcome of delivery, single liveborn: Other current conditions classifiable elsewhere of mother, delivered, with or without mention of antepartum condition; High vaginal laceration, delivered, with or without mention of antepartum condition; Carrier or suspected carrier of group B streptococcus</td>
<td>R2: Attending Phys/Prov → NMSP APRN (an advanced practice nurse) → Attending Phys/Prov [long duration, large variance]</td>
</tr>
</tbody>
</table>

Figure 5. Average duration and variance of each of the 25 topics.

Case Study of a Complex Patient

The long duration and variance of topics may be due to the complexity of patients. To illustrate we conducted a case study of a patient who is characterized by seven topics.

Figure 6 shows the distribution of the number of topics needed to characterize a patient. It can be seen that most of the patients are characterized by 1 or 2 topics. We say patients who are associated with multiple topics are complex patients. Different topics characterize different types of patients. If a patient is associated with multiple types of topics, his condition should be quite complex.

To illustrate, let us consider one patient who was characterized by 5 topics and is associated with the conditions of multiple myeloma, diabetes mellitus, esophageal reflux, urinary tract infection, and personal history of malignant neoplasm of breast. This patient is associated with multiple blocks coming from different topics, of which we show two examples in Table 5. The first block is B3: [Rehab Speech → Rehab Speech] which has a long duration and large variance, while the other is B4: [Medical Records – Scanner → NMH Physician-CPOE], which has a short duration and variance.

Block B3 is associated with 3 reasons, and block B4 is associated with 5 reasons as shown in Table 5. The difference between reason R2 and R3 of the block B3 is Med Rec Release of Info and Med Rec Quality. The reason R3 associated with Med Rec Quality requires more time. If the reason following Med Rec Quality is Attending Phys/Prov (R1) instead of Patient Care (R3), then the time duration will increase. For block B4, the associated five reasons require a shorter duration, which suggests the block is stable and efficient.
system can enable an HCO to investigate and refine inefficient and unstable workflows.

We note that this is a pilot study on workflow extraction and modeling, but that the findings are promising as a roadmap towards the future. Specially, we believe that this type of investigation can be enhanced through several lines of research.

First, the reasons behind inefficient workflows clearly need to be studied in a more refined manner. Our work characterizes inefficiencies as a function of complexity in the patients and the surrounding workflows. For instance, one of our examples illustrates that residents assisting attending physicians (via access charts) may be a possible cause of inefficient workflows. Yet this explanation is limited in its explanatory power. Alternatively, or perhaps additionally, the reasons for inefficient workflows may include varying clinical urgency, variation in resource availability, varied clinical and system experiences of users, or even the design of EHR system itself. Further investigation of these factors will likely yield additional workflow representation optimization opportunities.

Second, the learned 25 topics need to be confirmed by clinicians. To do so, we will need to provide more nuanced contextual information about the patients associated with these workflows. Currently, the context of these patients is limited to the ICD-9 codes billed to insurance companies, which certainly does not cover the detailed conditions of a patient.

Third, the workflows our technique inferred are not associated with specific diseases. This will make it difficult for HCOs to determine where to invest in workflow optimization. We plan to construct the association study between workflows and diseases through data mining and machine learning technologies as a next step.

**Conclusion**

Modeling workflows for healthcare is challenging due to the complexity of clinical processes. In this work, we introduced a framework to model workflows at multiple levels of granularity. We illustrated that this framework can enable the categorization of workflows into four classes based on their duration: i) stable efficient, ii) unstable efficient, iii) stable inefficient and iv) unstable inefficient. We performed an extensive evaluation on Northwestern EHR event logs in the inpatient setting, where the results showed almost 95% of blocks are stable as well as efficient, and that over 80% patients are associated with efficient workflows. We further provided several illustrations of the reasons for inefficiency in a workflow and posited the main reason may derive from complexity of patients and the fact that Northwestern is a teaching hospital where residents are trained. Nonetheless, the reasons for inefficiency of workflows are diverse and will require additional contextual information (e.g., nuanced clinical data on patients, resource allocations of HCOs, and experiences of care providers) to further investigate and optimize workflows accordingly.

**Table 5.** Two typical blocks and their corresponding reasons and duration.

<table>
<thead>
<tr>
<th>Block</th>
<th>Reason</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1: Med Rec Quality → Attending Phys/Prov (the primary physician responsible for an inpatient’s care)</td>
<td>440</td>
<td>450</td>
</tr>
<tr>
<td>R2: Med Rec Release of Info → Patient Care ( associated with nursing roles)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R3: Med Rec Quality → Patient Care</td>
<td>460</td>
<td>10</td>
</tr>
<tr>
<td>R4: Rehab Services Clinician → Rehab Quality Audit</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R5: Rehab Assigned Therapist → Rehab Services Clinician</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Discussion**

Clinical workflow modeling can be a challenging endeavor because of the complexity of patients, variability in the healthcare environment, and changeover in staff. Poor documentation of workflow can limit the adoption, or successful implementation, of EHR systems. Our work provides a framework to generate such workflows at multiple levels of granularity via data mining. Our evaluation of the duration of such workflows enables the clear partitioning of existing workflows into four types, depending on their length and variability. We believe that such a characterization of the
List of Abbreviations Used in this Paper: Rehab PT: Rehabilitation Physical Therapist; NMPG MD: A physician who belongs to the Northwestern Memorial Physicians Group; CPOE: Computerized Physician Order Entry; APRN: Advanced Practice Nurse Provider; SN-OR RN SC: Surgical Nurse Operating Room Service Coordinator (includes scheduling and patient experience issues); Radiology Mgr/RC: Radiology Manager Resource Coordinator (includes scheduling and patient experience issues); RAD - Mgr/RC: Radiology Manager and Resource Coordinator (includes scheduling and patient experience issues).

Acknowledgements: This work was supported in part by NIH grants K99LM011933 and R01LM010207, and NSF CNS grants 09-64392 and CNS 13-30491.

References

Jeeyae Choi, PhD, RN\(^1\), Kay Jansen, DNP, RN\(^1\), Amy Coenen PhD, RN, FAAN\(^1\)

\(^1\)University of Wisconsin-Milwaukee, College of Nursing, Milwaukee, WI, USA

Abstract

In recent years, Decision Support Systems (DSSs) have been developed and used to achieve “meaningful use”. One approach to developing DSSs is to translate clinical guidelines into a computer-interpretable format. However, there is no specific guideline modeling approach to translate nursing guidelines to computer-interpretable guidelines. This results in limited use of DSSs in nursing. Unified modeling language (UML) is a software writing language known to accurately represent the end-users’ perspective, due to its expressive characteristics. Furthermore, standard terminology enabled DSSs have been shown to smoothly integrate into existing health information systems. In order to facilitate development of nursing DSSs, the UML was used to represent a guideline for medication management for older adults encode with the International Classification for Nursing Practice (ICNP®). The UML was found to be a useful and sufficient tool to model a nursing guideline for a DSS.

Introduction

Clinical practice guidelines are indispensable to healthcare providers. Guidelines provide care recommendations which are synthesized from a systematic review of evidence to optimize patient care. Clinical practice guidelines are a vehicle to translate health care knowledge and outcomes of research into a user-friendly format. Many guidelines have been published by professional organizations, government agencies and health care institutions since they have been shown to reduce variation in practice and to improve the quality of care. Studies indicate that clinicians’ guideline compliance is increased when guidelines are implemented in a computerized decision support system (DSS) because they are able to deliver patient-specific care at the point of care\(^1,2\).

In recent years, meaningful use of health information technology has the expectation of improvement in the quality, safety, and even the cost of health care. Specific objectives to achieve meaningful use are developed by the Center for Medicare and Medicaid Services (CMS); one of the objectives involves using DSSs. CMS provides financial incentives to institutions that meet meaningful use requirements. To be eligible for incentive payments, one DSS should be incorporated into an institution’s electronic health record to support clinicians’ decision-making process\(^3\).

In spite of such important benefits, limited DSSs are used in nursing. One plausible reason is that DSSs developed for nurses often did not meet nurses’ information needs\(^4,5\). When a DSS is developed without accurate knowledge translated from a guideline, it does not provide accurate information to clinicians at the point of care. Studies show that formalization of guidelines as computer-interpretable guidelines (CIGs) makes a DSS more accurate and reliable\(^6,7\). Since most nursing guidelines are written in free text, tables and flow charts, it is necessary to develop CIGs to represent the underlying knowledge with explicit notations to define the semantics and provide reasoning mechanisms\(^8\). CIGs could facilitate development of various DSSs for nurses’ use, however, no formal modeling language to translate guidelines to CIGs has been developed for nursing use.

Unified Modeling Language (UML) is a software writing tool that has been widely used to model health information systems because of its useful effects of visualizing, specifying, constructing, and documenting the artifacts of a system from the end-users’ perspective\(^9\). It may also be a useful tool to model a nursing guideline from the nurses’ point of view.

DSSs also need to be integrated with existing health information systems for data reuse and patient-specific recommendations. Studies indicate that standard terminology enabled DSS can be seamlessly integrated with existing health information systems\(^10,11\). The purpose of this study is to illustrate steps involved in modeling a nursing guideline using the International Classification for Nursing Practice (ICNP®) and UML to develop a web-based medication management decision support system for older adults (Web-based medication management DSS).
Background

Computer-interpretable guidelines (CIGs)

A CIG is a systematically visualized and specified guideline by graphic notations to describe core knowledge of a guideline. This translation of a guideline has been useful to develop DSSs because it allows accurate and complete interpretation of guideline content. Studies show that CIG-based DSSs are easy to maintain (i.e. upgrading) and are adapted easily in local health care settings. Accordingly, many modeling approaches have been developed to represent and execute guidelines. Three distinct modeling approaches are found in the medical informatics community and categorized as: 1) Document models (i.e. Guideline Elements Model: GEM), 2) Decision tree and probabilistic models (i.e. Fuzzy Cognitive Maps: FCM) and 3) Task-network models (i.e. Guideline Interchange Format: GLIF). Despite such efforts, only a few guideline modeling approaches have progressed beyond the prototype stage.

Unified Modeling Language (UML)

The UML is a standard software writing language that was used to design and develop software systems. It is also an expressive language that can describe all the viewpoints needed to develop systems. It is used to: 1) visualize, 2) specify, 3) construct, and 4) document the elements of a software system. UML facilitates communication between developers and end-users of software by allowing them to examine objects in a system. Its ability to specify allows the UML to build a precise, unambiguous and complete system. A model constructed by the UML can also be mapped directly to an object-oriented programming language (i.e. Java or C++). Thus, a system prototype can be tested in early stages of the development, allowing for early refinement. In addition, UML can be used to document whole processes of developmental stages and tasks. These records help control, measure, and maintain a system after its deployment.

Standardized terminology in decision support systems

Due to the government emphasis on “meaningful use” for Electronic Health Records, DSSs are widely introduced in healthcare institutions. However, their adoption by clinicians has been slow and many challenges have emerged in development and implementation stages. One of the challenges is incorporating standardized terminology in a DSS. According to Ahmadian and colleagues’ literature review and survey, 58% of DSS developers experienced problems at developmental stages and 92% of these problems were related to data standardization. In addition to this, terminology standardization is an important requirement for scalable DSSs. Standard terminology and information models to represent health care data should be incorporated in DSSs for seamless integration.

To increase adoption and interoperability potential, a standard nursing terminology, ICNP was selected because it is a compositional terminology that facilitates the development of and the cross-mapping among local terms and existing terminologies. ICNP includes nursing diagnoses, nursing interventions and nursing-sensitive patient outcomes that describe nursing practice. The use of a standard nursing terminology such as ICNP facilitates description of nursing diagnoses, interventions, and outcomes for medication management nursing care and extraction of these data for later use.

Older adults and medication management

For older adults, the desire to remain independent as long as possible is critical. Research suggests that the ability to remain independent is highly dependent on the ability to manage medications. Multiple health conditions in the aging population often result in complex medication regimens that make this population vulnerable to adverse reactions to medications. The need for an electronic evidence-based decision support system to assist nurses in managing medications for older adults living the community is imperative. This type of computer system also has the potential to improve patients’ self-management of their health.

In recent years, health information technology has been used to improve care for older adults. Automated assessment tools and DSSs were developed and embedded in electronic health records. Nursing studies regarding teleHealth interventions were conducted. It is the appropriate time to develop a nursing DSS for older adults living in the community.
Modeling Steps

Step 1: Selection of a guideline for medication management for older adults

In the United State, approximately three million older adults are admitted to nursing homes because of their inability of manage a complicated medication regimen. It is estimated that the annual care cost for this population is $177 billion dollars17,21. Approximately 30% of hospitalizations of older adults are drug related and older adults who are discharged from the hospital with more than five medications are more likely to visit the emergency department and to be re-hospitalized within six months22. Unfortunately, due to complications with medications, older adults may not be able to live independently in the community. Nurses, aided by evidence-based decision support can partner with older adults and families in the community to safely manage medication.

A nursing practice guideline, “Medication Management of the Community-Dwelling Older Adult” (MM-guideline) was selected as a content source for this study. This evidence-based guideline was developed through support from the US Agency for Healthcare Research and Quality23.

Step 2: Extraction of concepts related to nursing processes in a selected guideline

After reviewing the MM-guideline, concepts related to medication management nursing processes and associated elements of each process were extracted. Processes and their associated elements were tabulated in a table. For example, “Risk assessment” was identified as a nursing process with ten associated elements. A total of eight processes were extracted and validated by a nurse expert. Table 1 shows four examples of extracted nursing processes and their associated elements.

Table 1. Four examples of nursing processes and associated elements extracted from a guideline

<table>
<thead>
<tr>
<th>Processes</th>
<th>Associated elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk assessment</td>
<td>name, age, living condition, chronic diseases, psychiatric diagnoses, number of medication, medication dosing schedule, complex medication, number of prescribing providers, changes in medication regimen within 1 month</td>
</tr>
<tr>
<td>Medication reconciliation</td>
<td>medication list, associated diagnosis, prescriber, dose, frequency, Beers criteria, adverse drug reaction, potential drug-drug reactions</td>
</tr>
<tr>
<td>Medication procurement</td>
<td>difficulty obtaining or refilling prescriptions, viewing medications as too costly, funds to pay for medications or other financial burden, pharmacy delivery and/or refill reminder</td>
</tr>
<tr>
<td>Care plan</td>
<td>problems, interventions, expected outcomes</td>
</tr>
</tbody>
</table>

Step 3: Constructing an activity diagram

UML activity diagrams are typically used for computational and organizational processes modeling, for modeling the logic captured by use cases or scenarios, or for modeling the detailed logic of a processing rule. The activity diagram focuses on the action sequence of execution and the conditions that trigger or guard those actions. Activity diagrams are used for object-oriented software development while flow charts and data flow diagrams are used for structured software development24.

In order to create an activity diagram, actions and decisions are needed. All nursing processes in step 2 were identified as essential and necessary actions to manage medication for older adults. Then, any decision or parallel decisions which needed to be made were identified. Starting from an initial state symbol, actions were laid out in an order, and the flow from one action to another was represented using an arrow. The diagram ended with a final state symbol (Figure 1). Upon completion, a nurse expert, who specialized in community nursing with more than 10 years of research experience, validated the logical consistency of the activity diagram. Necessary revisions were made based on the nurse expert’s feedback. For instance, an assessment of each area of risk such as “Medication reconciliation” and “Medication procurement” was skipped algorithmically when there was no nursing problem identified in the first activity diagram. The nurse expert recommended that a nurse needed to assess all areas of risk
even if there were no problems identified in the initial risk assessment. In the second activity diagram, this suggestion was incorporated.

**Step 4: Construction of algorithms and mapping with ICNP**

Algorithms are well-defined procedures and an essential part of any computer system\(^\text{25}\). They can be written in pseudocode in early software developmental stages to test logics. Pseudocode is simple English like language that allows basic entity relationships to be expressed in a concise format. It can be parsed to build software modules and be rendered to computer languages (i.e. Java, C++, or Ruby programming language, etc.).

After algorithms were constructed based on nursing processes and logics from step 2 and 3, pseudocodes for each nursing process were written. Figure 2 shows examples of pseudocodes written for a web-based medication management DSS. Total seven care plans were extracted from the MM-guideline and data items of problem, intervention and expected outcomes of all seven care plans were mapped with ICNP.

![Diagram](image)

**Figure 1.** An activity diagram constructed for a web-based medication management DSS

**Step 5: Identifying software requirement specification**

A software requirements specification is a comprehensive description of software under development. It describes what the software will do, how it will interact with end users, what the boundary of the software is and how it will be expected to perform. Purpose of this description is to minimizes the time, effort and cost required by developers to achieve desired goals for software users. In this study, guidelines from IEEE\(^\text{26}\) were used to describe the web-based medication management DSS.
For example, eight required functionalities were identified and written as:

1. Web-based medication management DSS shall display data entry screens for user.
2. Web-based medication management DSS shall allow user to select a patient.
3. Web-based medication management DSS shall allow user to enter new patient.
4. Web-based medication management DSS shall display screens with data entry tools such as checkboxes, dropdown menu, or text-field.
5. Web-based medication management DSS shall allow user to move screens forward and backward.
6. Web-based medication management DSS shall display nursing intervention list.
7. Web-based medication management DSS shall allow user select interventions from the list.
8. Web-based medication management DSS shall allow user to save identified risk factors and provided nursing interventions into database.

**Figure 2.** Examples of algorithms constructed for a web-based medication management DSS

---

Step 6: A web-based medication management DSS

A web-based medication management DSS was programmed based on information from steps 1 to 5. Only two instances of clarification between team members and a programmer were needed. Once a prototype was programmed, team members performed alpha testing for a web-based medication management DSS. Three functions of a web-based system were not working correctly and were adjusted during alpha testing. Figure 3 shows the initial (middle), risk assessment (left lower) and Medication reconciliation assessment (right lower) screens of a developed system.
Discussion

Unified Modeling Language (UML) was developed to compose software and to assist the developer to model a computer system. Its descriptive nature describes a potential computer system accurately before implementation. For this reason, the UML was selected as the tool for modeling a guideline in this study. During the modeling processes, the UML was found to be a useful and sufficient tool to describe the Medication Management of the Community-Dwelling Older Adult guideline for a web-based system. It may be recommended that nurse informaticians utilize the UML when they model a nursing guideline for a decision support system for nurses’ use.

The UML has been used for modeling in many studies, but only one study has been reported in the nursing domain. Therefore, representing this nursing guideline with the UML is a significant first step in preparing nursing guidelines for implementation as decision support systems. Since there are limited computer systems for nurses’ use to date, the approach described may be useful in two ways: 1) it may facilitate interactions between a nursing informatician (on behalf of nurse users) and a developer to implement a computer system that nurses prefer; and 2) it may enable the maintenance of a model of a computer system that is easily modified for upgrades.

Limitations

The study findings have limited generalizability because the 1) UML modeling process was tested with only one guideline, 2) accuracy and completeness of translation of UML should be compared with those of other guideline modeling approaches (i.e. GLIF) and 3) the web-based medication management system was not tested with real patients’ data. Future research is necessary to develop decision support systems by following steps in this study with different guidelines and to test this system in community or public health nursing practice.

Acknowledgments

Authors acknowledge support for this study by the University of Wisconsin College of Nursing Harriet H. Werely Award and the International Council of Nurses.
References


LORD: a phenotype-genotype semantically integrated biomedical data tool to support rare disease diagnosis coding in health information systems

Remy Choquet, PhD1,2, Meriem Maaroufi1,2, Yannick Fonjallaz3, Albane de Carrara1, Pierre-Yves Vandenbussche2, Ferdinand Dhombres*, MD, PhD2,3, Paul Landais*, MD, PhD3,4

1Banque Nationale de Données Maladies Rares, Hôpital Necker Enfants Malades, APHP, F-75015, Paris, France; 2INSERM, U1142, LIMICS, F-75006, Paris, France; 3Service de Médecine Fœtale, Hôpital Armand Trousseau, APHP, UPMC, Paris, France; 4Montpellier1 University, EA2415 & BESPIM, University Hospital, Nîmes, France; *both authors equally contributed to the work

Abstract

Characterizing a rare disease diagnosis for a given patient is often made through expert’s networks. It is a complex task that could evolve over time depending on the natural history of the disease and the evolution of the scientific knowledge. Most rare diseases have genetic causes and recent improvements of sequencing techniques contribute to the discovery of many new diseases every year. Diagnosis coding in the rare disease field requires data from multiple knowledge bases to be aggregated in order to offer the clinician a global information space from possible diagnosis to clinical signs (phenotypes) and known genetic mutations (genotype). Nowadays, the major barrier to the coding activity is the lack of consolidation of such information scattered in different thesaurus such as Orphanet, OMIM or HPO. The Linking Open data for Rare Diseases (LORD) web portal we developed stands as the first attempt to fill this gap by offering an integrated view of 8,400 rare diseases linked to more than 14,500 signs and 3,270 genes. The application provides a browsing feature to navigate through the relationships between diseases, signs and genes, and some Application Programming Interfaces to help its integration in health information systems in routine.

Introduction

Eighty percent of Rare Diseases (RD) have a genetic origin. Presently, 3,000 among more than 7,000 rare diseases have an identified gene.1 Moreover, clinical signs are often used to characterize RD.2 Even when a consensual disease classification is used in a registry for RD centers,3 the “not confirmed” category represents 1 out of 4 cases (17% undetermined, 9% could not be classified yet). The identification of patients with rare diseases at international level is a key requirement to accelerate patient recruitment for clinical trials and cohorts. This work is proposed within the framework of the second French National Plan for Rare Diseases (2011-2016) where the French National Database for Rare Diseases infrastructure (BNDMR)3 is defined. The BNDMR is a priority project of the National Plan and aims at gathering a common minimal data set for all rare diseases patients at nationwide scale5 to facilitate patient recruitment in clinical trials and in cohorts as well as public health studies. A common terminology is required to register patient diagnosis and enable statistical use of the data. The French Ministry of Health encourages the use of Orphanet6 for national coding of RD while the European Commission recently decided to recommend Orphanet as an European terminology for RD diagnosis coding.7

In order to properly code rare disease patients into health information systems (medical records, registries or cohorts), several issues have been discussed by health information specialists and clinicians: (i) poor information on rare disease is available in standard morbidity classifications (ICD-9 to ICD-10); the ICD classification is often used for reimbursement purposes in pay-for-performance systems, and is not well adapted to describe patient’s clinical status, (ii) to code a patient diagnosis, there is a possible need to gather information from other knowledge bases, such as genetic and phenotype information, (iii) navigating within more than 7,000 RD cannot be done through a tree navigation or a list navigation. Search algorithms are nice but context dependent navigation is a key feature (a context can be a medical field), (iv) RD are often multi-systemic, thus mono-parental classifications such as ICD might not be adapted.

Coding systems (thesaurus, classifications, terminologies, controlled vocabularies) are used to enable controlled data entry in health information systems in order to facilitate data analysis for quality of care,8 epidemiology9 or research.10 Historically, the International Classification of Diseases (ICD) was used to measure mortality in
populations. Its usage was progressively redefined to measure morbidity and was then integrated within global hospital information systems with some difficulties. In the rare disease field this classification does not provide us with the necessary RD completeness (only 300 rare diseases are described) and when described, rare diseases are not necessarily used as the primary or secondary diagnosis since the visit of the patient might be related to a comorbidity (e.g. respiratory distress for a Cystic fibrosis patient). There are few resources available that can help in classifying RD patients: Orphanet, OMIM, HPO, the GARD disease nomenclature (https://rarediseases.info.nih.gov/gard) or SNOMED CT. These resources do not present the same information type or completeness but they could be used together to specify patient rare conditions with a relatively good phenotypic or genotypic precision.

In this paper, we will not discuss the adequacy of using Orphanet as a coding system to record morbidity in the rare diseases field nor the curation made by Orphanet, OMIM and HPO to interlink their resources. We will focus on implementing a tool that could use the Orphanet proposed views (multi-parental and multi-classified diseases) to help guiding professionals to the right disease (medical coding assistance). Besides, we will propose to enrich each of the Orphanet disease with genetic and phenotypic information available from OMIM and HPO to provide a richer presentation of the disease. Many terminology sources are available, and it is now crucial to be able to bring data from different sources together into one application to help clinicians in their coding task. Available web applications do provide parts of the information for a given disease (http://www.orpha.net, http://omim.org). Few projects offer to link the diseases all together (Diseasecard, HeTOP, or Orpha.net) but they do not provide the actual information from all sources into the same web interface. We propose in this paper a new generation application for rare diseases coding assistance based on semantic web/RDF technologies for the rich semantic linked data DB on one hand and NoSQL/JSON for the web services technologies access on the other hand.

Methods

The semantic integration of biomedical knowledge sources and datasets is a complex data engineering issue. In order to enable the integration of different datasets, our architecture is based on a first semantic integration layer that deals with the heterogeneity of source formats as well as the semantic heterogeneity of proposed representations of diseases. We then defined a presentation layer composed of a NoSQL database well fit to represent JSON complex biomedical objects to be displayed through a Ruby on Rails web application. That architecture combines the use of semantic web technologies that have great expressivity capabilities (RDF, OWL) but have a footprint on performance for real-time applications, with the flexibility and performance of NoSQL technologies to provide the best user experience when navigating through complex biomedical data.

---

**Figure 1.** Genetic-phenotypic semantic database integration architecture for rare diseases. The first layer function deals with the technical and the semantic integration of various data sources. The second layer deals with presenting the data and navigating through views.

1. **Integration layer**

The integration of various externally curated multi-lingual biomedical datasets requires expressive models and semantic robustness to express exact, broader-than and narrower-than alignments as well as reasoning. We have
chosen to integrate the datasets into a RDF triple store (Openlink Virtuoso®) through a homemade integration layer (xml and java data integrators). Each data integrator was developed specifically for each data source as data structures and format of sources were heterogeneous. Orphanet data is released using xml based on a local model extraction without xsd, OMIM is available as a set of flat text files and HPO as an OBO or OWL file. Whilst the HPO file has a rich semantic web compatible serialization format, Orphanet and OMIM are not semantic web compatible, semantic links between concepts were thus rebuilt to be compliant with RDF. Table 1 reflects the volume of integrated triples in the RDF triple store.

Table 1. Dataset integrated triple volumes (at 1st January 2015)

<table>
<thead>
<tr>
<th>Categories</th>
<th>Information</th>
<th>Orphanet</th>
<th>OMIM</th>
<th>HPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disorder</td>
<td>PrefLabels</td>
<td>8,489</td>
<td>7,666</td>
<td>4,170</td>
</tr>
<tr>
<td></td>
<td>Synonyms</td>
<td>10,412</td>
<td>8,650</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>472</td>
<td>2,604</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Diagnosis</td>
<td>0</td>
<td>412</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Clinical features</td>
<td>0</td>
<td>4,783</td>
<td>0</td>
</tr>
<tr>
<td>Signs</td>
<td>PrefLabels</td>
<td>1,360</td>
<td>0</td>
<td>12,626</td>
</tr>
<tr>
<td></td>
<td>Synonyms</td>
<td>0</td>
<td>0</td>
<td>6,424</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>0</td>
<td>0</td>
<td>6,844</td>
</tr>
<tr>
<td></td>
<td>Broaders</td>
<td>1,310</td>
<td>0</td>
<td>13,371</td>
</tr>
<tr>
<td>Genes</td>
<td>PrefLabels</td>
<td>3,270</td>
<td>14,419</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Synonyms</td>
<td>7,168</td>
<td>21,275</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Symbol</td>
<td>3,270</td>
<td>23,429</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td>0</td>
<td>20,639</td>
<td>0</td>
</tr>
</tbody>
</table>

Overlaps between these three datasets to represent disorders (diseases) were enumerated in the disease ontology. The Orphanet source also offers cross referencing to HPO and OMIM and other datasets. They also produce more than 30 views (named diseases classifications by Orphanet) organized per rare disease groups (including 1 non-rare diseases group). Each disease group is organized into several entities such as categories of diseases, diseases, clinical subtypes, etiological subtypes, etc. These views are curated by Orphanet and will be used to implement the navigation functions of the LORD application. Rare diseases are numerous and can often have several medical expressions. For instance, patients diagnosed with Fabry disease [17] can present several anomalies: renal, cardiac, neurologic, dermatologic, ophthalmic etc. The classification can also be used to make differential diagnosis, if the patient present a Cataract associated with a metabolic disease [18], then the patient could be suffering from Fabry disease, but also Refsum disease, De Bary syndrome, Galactosemia or Alpha-mannosidosis diseases (given by Orphanet views). Searching from clinical signs is not supported in this version of the LORD application.

2. Presentation layer

Semantically enriched biomedical data stored in native RDF triplestore generally suffers from a performance footprint when queried on a large scale application with real-time navigation features on the web. On the other hand, new generation web applications (Ruby on Rails, Backbone.js) can be easily built up using a full web service architecture managing JSON objects. JSON objects are similar to rich XML objects and can be stored in NoSQL databases such as MongoDB, allowing flexibility for model changes as well as best in class performance and scalability using sharding techniques if needed (data atomization over MongoDB servers and query balancing).

To generate JSON objects, SPARQL queries are processed on the RDF triplestore, and resulted JSON objects are stored into the NoSQL database in batch mode, every day. Such a process can be applied from the same RDF triplestore to numerous presentation databases for various projects if needed. Once integrated and formatted for presentation as JSON objects, the NoSQL database is indexed and searched through an Elasticsearch engine (http://www.elasticsearch.org) enabling auto-completion, multi-field, ranked searches as well as various possible optimizations. The search engine enables search over disorder names and synonyms, Orphanet numbers and genes. Given the nature of the Orphanet resource (multi-view and multi-parental), we pre-calculated all possible concept parents and sons to help with the navigation feature of the interactive graphical interface. Hence, when a user selects a disorder, he can view direct neighbors of the given disease.
The result is returned and displayed to the user in a graphical and textual interface. We defined a navigational area and a content area. Users can navigate through classifications, filter classifications and diseases, and add current disease to a disease ‘basket’ for exportation if the application is integrated with an EHR or a patient registration system. They have then access to all information available from Orphanet, HPO and OMIM for the selected disease on the same screen. The Human Rare Disease core Ontology (HRDO) provides a meta-model for this presentation layer implementation.

Figure 2. The disease interface of the LORD application. The filtering view gives access to filters based on medical specialities. The graphical tree navigation enables navigation through diseases, symptoms and group of diseases. The external thesaurus links allow the user to be redirected to source sites pages of diagnosis. The disease content view represents disease definition data from Orphanet, signs from HPO and clinical synopsis and genetic data from OMIM.

Results

One of the objectives of the project was to be able to create a common, curated and semantically rich knowledge repository to help managing the complexity and heterogeneity of biomedical databases. Data are openly available but with poor usability, as they are not integrated into a single application. Update frequencies are heterogeneous and as the sources are interlinking their datasets, they present some data in common. It is therefore necessary to build an integrated semantic database that can also serve as a national terminology server for several purposes: search terms, search based on clinical signs, disease complete data retrieval, integration within another application of the curated database. The RDF datastore can be queried over the semantic web SPARQL query language which is the most advanced query language enabling reasoning, inference and federation. It is compatible with Linked Data principles. Data can also be queried directly through a secured web service directly on the NoSQL database, depending on the use case of the user.
When connecting to the application (http://enlord.bndmr.fr) the user can browse the data through 2 modes. Either using the search function, or by browsing using proposed Orphanet views. An Elasticsearch engine that provides multi-criteria searches over several data sources powers the search function. Users can search through diseases preferred label, alternative label, gene or Orphanet number. The search engine enables word search regardless of the position of the word within the disease label. Given the length of some disease names, this feature is important. For instance, the *Anhidrotic ectodermal dysplasia - immunodeficiency - osteopetrosis – lymphedema* Orphanet disease (http://enlord.bndmr.fr/#disorders/69088) which is composed as a set of symptoms. By navigating backwards in the hierarchy to *Osteopetrosis* (http://enlord.bndmr.fr/#disorders/2781) the user can search over a limited set of diseases that share Osteopetrosis, if it is the observed symptom.

When users are navigating through diseases graphs, they can filter the displayed concepts. For example, while looking at *Cystinosis* information (http://enlord.bndmr.fr/#disorders/213/97966), the user can filter on Rare eye diseases (if he is an Ophthalmologist) and then navigate through symptoms that are related to his medical expertise: *Unclassified maculopathy, metabolic disease with corneal opacity or metabolic disease with pigmentary retinisis*. We should emphasis that the nature of the disease “groups” here differs from the previous example. While the first “groups” are related to symptoms of a disease, the seconds are diseases classifiers.

The user then can have access to textual information from OMIM and HPO for a given disease and epidemiological information from Orphanet. As relations between data sources can be one to many (1 Orphanet disease for n OMIM entries) then we have developed an OMIM entry selector to help user navigation and data retrieval. From the web service standpoint, all data of all OMIM entries are gathered in the same JSON object for an individual Orphanet disease.

**Discussion**

We defined and built a new generation web application to help clinicians and health information specialists to navigate in the Orphanet rare diseases for diagnosis coding. We first integrated and curated various terminologies and classifications (Orphanet, HPO, OMIM, HRDO) related to our domain within a semantic database adapted to reasoning and compatible with linked data principles. Once curated and qualified, the semantic biomedical data is transformed into JSON complex objects and stored in a NoSQL database for presentation and navigation through a new generation web portal. This 2-layer application architecture has proven to combine both generally distinct paradigms: expressivity and performance.

The biomedical data we integrated can be used in other applications, either using directly SPARQL (SPARQL endpoint not publicly available yet), or through web services. For instance, through the search engine, the “cystin” search query could be executed following a command line or web request (GET) http://enlord.bndmr.fr/disorders/search/cystin.json?[page=N]. The result will be a JSON object:

{"term":"cystin","page":1,"total":1,"lastPage":1,"results":[Full collection of diseases found]} that can be directly processed within a third party application. The access to the data through APIs and the application browser are freely available in French and English.

The application was evaluated by clinicians and information managers and used in real setting. The quantity of information on screen was judged good as well as the navigation functions. Clinical geneticists requested more OMIM data and clinicians less Orphanet concepts and more precision upon signs. Some indicated that not all information is useful to all professionals. Medical experts also argued that Orphanet classifications cannot be seen as a diagnosis assistance tool (guiding the clinician from presentations of the disease to the precise disease). Others considered that Orphanet as a master terminology might not be sufficient, as rare manifestations do not have necessary a known diagnosis and diagnosis certainty is a key requirement in the rare disease field. Geneticists also criticized data update latency of OMIM and Orphanet as new genes or mutations are found regularly. Last and most important, not all patients can be given a diagnosis at a certain point of time. The disease might naturally evolve (in case of a genetic disease) over time or the patient might or might not express it. Hence the application should update the thesaurus rapidly but, even with fast updates, the application solely relies on pre-curated terminologies by external parties. The presence of a correct entity to represent at the right granularity (clinical sub-type, disease, group of diseases, disease category) is, for the moment, entirely relying on our data providers. Granularity is also seen by health professionals as a barrier to a homogeneous national data collection for later statistical exploitation of data. A better updating process of these databases by users should be set in order to help Orphanet/HPO/OMIM in their curation tasks which are time consuming. Finally, for the coding task, instructions for coding RD should be elaborated in collaboration with RD expert networks. We foresee in the next months a formal evaluation of
the application whilst the new generation application for data collection for RD centres is rolled out in France (BaMaRa). During the rollout, data accuracy and quality, as well as navigation within the LORD application will be evaluated. Also, coding cases will be set to assess the views proposed by Orphanet as well as the navigational behaviours of the application for that particular purpose.

The application is not yet used in routine in France in all hospital information systems as the coding of rare diseases has not yet started in hospital information systems. As at today, the application is used by 200 users monthly. Users are located in France, Spain, Belgium, Switzerland, Germany and the United States. It is also used by some French university hospitals in clinical routine and by Orphanet France and Switzerland for navigating through their data and for curation.

This first version of the application could be enhanced in many ways. First, we could integrate other existing initiatives like DiseaseCard, which offers more links to external sources than Orphanet. Second, we could help the coding specialist or the expert physician to code the diagnosis from the patient discharge letters. Third, we could develop a search by sign engine that could provide a list of candidate diseases. Fourth, we could propose on the same infrastructure a signs navigation engine for HPO. Fifth, we could enrich the coding possibilities to other nomenclatures such as HPO, OMIM, CIM10, SNOMED since not all rare diseases are represented in Orphanet only. The application is also seen as a potential coding application for the RD-ACTION European joint action (2015-2018) that could also serve as a European terminology service integrated hub for coding rare diseases across European countries.

Conclusion

We introduced a new generation application to lighten the burden of rare diseases codification of cases based on controlled vocabularies. LORD results are significant as (1) the proposed architecture is set on the latest standards (Semantic Web, RDF, JSON, NoSQL, REST) and is flexible enough to curate complex biomedical objects and navigate through them easily at large scale, (2) it brings to users many information about a rare disease into the same UI (genotypic, phenotypic, epidemiological) (3) it could help in some rare cases, in defining the right disease by disease similarity analysis for differential diagnosis, (4) it can help in being more precise in the classification of a disease, (5) it facilitates the specialist navigation through complex graphs of concepts by applying filters on the data (rare eye diseases, rare kidney diseases, etc.).

Acknowledgements

The French Ministry of Health funded this work under the second national plan for rare diseases. We thank our data sources (Orphanet, OMIM, HPO) for their great work in curating and interlinking constantly their respective databases. We warmly thank Céline Angin for her editing and design work.

References


Characterization of the Context of Drug Concepts in Research Protocols: An Empiric Study to Guide Ontology Development

James J. Cimino, MD¹,² and Vojtech Huser, MD, PhD²

¹Informatics Institute, University of Alabama at Birmingham, Birmingham, Alabama
²Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland

Abstract

We examined a large body of research study documents (protocols) to identify mentions of drug concepts and established base concepts and roles needed to characterize the semantics of these instances. We found these concepts in three general situations: background knowledge about the drug, study procedures involving the drug, and other roles of the drug in the study. We identified 18 more specific contexts (e.g., adverse event information, administration and dosing of the drug, and interactions between the study drug and other drugs). The ontology was validated against a test set of protocol documents from NIH and ClinicalTrial.gov. The goal is to support the automated extraction of drug information from protocol documents to support functions such as study retrieval, determination of subject eligibility, generation of order sets, and creation of logic for decision support alerts and reminders. Further work is needed to formally extend existing ontologies of clinical research.

Introduction

Clinical research studies involve complex work flow that might benefit greatly from automated processes to assist with tasks such as retrieval of study information, subject recruitment, data collection procedures, administration of study interventions, resource planning, and decision support for clinicians caring for research subjects. Such processes will necessarily depend on representing study concepts and procedural knowledge through formal mechanisms such as ontologies and logic. To that end, Sim and colleagues¹ have worked to develop the Ontology of Clinical Research (OCRe) intended to represent a broad range of research study concepts. Efforts to define and use such concepts have been especially successful with respect to characterizing the eligibility criteria by which research subjects are selected for or excluded from studies.²

Research protocols still typically consist of lengthy, loosely structured text documents—a format that confounds the application of automated tools. However, Tu and colleagues³ demonstrated that formal representations of study concepts could be developed by identifying interesting text phrases in protocol documents and then representing their semantics using a template-based formalism called ERGO. Their findings support the notion that, when combined with natural language processing, automated extraction of research concepts (such as eligibility criteria) and their conversion to computable form is possible. Drugs are particularly interesting because of their involvement with many aspects of research. For example, over 35% of the subject eligibility phrases selected for the ERGO study, reported by Ross and colleagues,⁴ involved drug concepts (Sim I, personal communication).

In work at the National Institutes of Health (NIH),⁵ we examined the roles of drugs in research studies from a somewhat different perspective, namely documenting the drugs that were restricted from use in research subjects. Like the ERGO study, we found many instances in which subjects would be excluded from a study due to previous exposure to a particular drug or class of drugs. However, we also found many restrictions that related to use of drugs during the course of the study, sometimes in complex situations (e.g., “do not give any drugs that will prolong the patient’s QT interval within 5 half-lives of the administration of the study drug”).

As a result of our analysis of NIH research studies, we developed a medical logic module for use in electronic health records that would warn clinicians attempting to order a drug that is restricted in one of the studies in which the patient at hand was participating. As part of an effort to expand decision support around “drug-study interactions”, we explored the relationships involving drugs in all aspects of research study documents (protocols) with the goal of identifying the breadth of contexts in which drug concepts appear. This empirical work provides one approach to extending existing ontologies so that they, in turn can be used represent drug information to automate a variety of research-related tasks, such as study retrieval, subject eligibility determination, generation of research-specific order sets, and creation of logic for decision support alerts and reminders. This paper reports on the results of a formal analysis of study documents related to research at the NIH Clinical Center and in the National Center for...
Biotechnology Information’s ClinicalTrials.gov database, establishment of a base set of roles, and speculates on the potential advantages of being able to determine which drugs have which roles in a research study.

Methods

Because research documents may discuss drug concepts without mentioning specific drugs (for example, by mentioning classes of drugs or by simply saying “drugs”), we developed a set of search words for identifying direct and indirect references to drug concepts. The initial set was based on text search words we used in our previous study, combined with key words in the text phrases identified in the ERGO study.

We obtained a convenience sample of protocol documents for active and terminated studies at the NIH Clinical Center, conducted by researchers at 13 different NIH institutes and centers. We set aside 10% of these documents for use as a test set and then searched the remaining documents (the development set) for ontology development. We first searched a subset of the development set to locate drug terms (e.g., specific named drugs and drug classes). We then examined occurrences of these particular drug terms in the development set to find additional words associated with the drug terms that might be useful in searching study documents.

We searched the development set, one study document at a time, for any occurrences of any of the search words (with stemming, such that “chemotherap” would find both “chemotherapy” and “chemotherapeutic”). One of us (JJC) reviewed each occurrence to determine if it involved information about a drug (for example, the search word “administer” might involve a drug or a survey or might simply refer to administration of the study itself).

Once a relevant phrase was identified, it was copied to a spreadsheet. This task was complicated by the fact that most of the documents were created using the Portable Document Format (PDF). While the optical character recognition function in Adobe Acrobat (Adobe Systems Incorporated, San Jose, CA) made scanned documents searchable, many typographical errors needed to be corrected manually. Subsequently encountered phrases that were identical or nearly identical were not included in the spreadsheet. The process was halted when ten consecutive protocols were searched without identifying any phrases that differed semantically from previous findings. Once the spreadsheet of phrases was assembled, ontology creation was accomplished by through manual inspection, classification and description, similar to methods used in other analyses of research study documents.

To the test set we added study descriptions (in XML format) obtained by searching ClinicalTrials.gov for the “top 100” studies that contained the term “2015”. This expanded test set was provided to one of us who had not seen the ontology (VH) with instructions to highlight 100 phrases that involved drugs in some way. The researcher was encouraged to ignore phrases that were similar to previously highlighted phrases to improve the variety, and thus the coverage, of the test phrases. Once the phrases were identified, they were coded independently by each of us using the ontology. Results were pooled and inter-rater discrepancies were reviewed and reconciled by both coders.

Results

Development of Source Material

The convenience sample included study documents from 866 active and terminated studies at the NIH Clinical Center, conducted by researchers at 13 different NIH institutes and centers. We set aside 87 (10%) of these documents for use as a test set and used the remaining 779 documents as the search set.

Our original set of three search words (“exclu”, “prohib”, and “concomitant”) was expanded with 54 words from the ERGO data. The search for 20 drug terms identified “5-FU”, “acetaminophen”, “Antibiotic”, “Antihistamine”, “antipyretic”, “aspirin”, “baclofen”, “Botulinum”, “cisplatin”, “Dexamethasone”, “doxorubicin”, “Filgastrim”, “ibuprofen”, “ISOVUE”, “Leucovorin”, “lithium”, “metoprolol”, “statin”, “vaccine”, and “vitamin”. We found these drug terms 893 times in 52 documents, which helped identify 12 additional search words (“intrathecal”, “continue”, “dosage”, “allowable”, “infusion”, “monotherapy”, “formulary”, “nonformulary”, “rescue”, “eluted”, “elution”, and “chemoembolization”). The final set of 69 search words is shown in Figure 1.

Searching the 779 protocol documents in the search set with the text words shown in Figure 1 yielded 481,077 results, the vast majority of which were false positives. We therefore started our review with the 103 documents from studies conducted by Clinical Center investigators and then continued the review with selected documents from recent studies conducted by investigators at each of the other 12 clinically active institutes before halting. Altogether, this review identified 72 examples of semantically distinct text samples.

* https://clinicaltrials.gov/ct2/resources/download
Ontology Creation

Semantic categorization began by noting that many drug-related phrases involved background information about the drug that was not specific to the study, for example, history of experience with the drug or pharmacologic information about the drug. We also noted that when drugs were discussed in relation to the study, they might be related to an actual procedure to be carried out in the study (for example, administration of the drug) or they might be related in some other way (for example, data collection about use of the drug). We initially classified each of the 72 text samples into one of three classes: Background Information, Study Procedure, and Relation of Drug to Study. Within the classes, we then further classified text samples as relating to specific background information (3), study procedures (7) or other relationships (8). The resulting three classes and 18 specific drug roles is shown in Table 1.

Table 1. Eighteen concepts characterizing the roles of drugs in research protocols.

<table>
<thead>
<tr>
<th>Class</th>
<th>Specific Role</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Background</td>
<td>A.1 Pharmacologic effect information</td>
<td>extra-arterially, acetylcholine produces a dose-dependant vasodilation that lasts only for a few minutes</td>
</tr>
<tr>
<td>information</td>
<td>A.2 Description of drug pharmacodynamics/pharmacokinetics</td>
<td>While dabigatran itself is not a substrate of P-glycoprotein (P-gp), its inactive pro-drug, dabigatran etexilate, is a substrate of P-gp. Co-administration of dabigatran etexilate with P-gp modulators has resulted in significant changes in dabigatran exposure.</td>
</tr>
<tr>
<td>information</td>
<td>A.3 Adverse event information</td>
<td>hetastarch given during granulocyte administration rarely causes mild allergic reaction</td>
</tr>
<tr>
<td>B. Study Procedure</td>
<td>B.1 Collection/retention of data about drug use</td>
<td>relevant medical information such as current medications, presence of any communicable diseases and allergy history will be documented</td>
</tr>
<tr>
<td></td>
<td>B.2 Administration and dosing of drug</td>
<td>The initial dose of rtPA will not exceed 10 mg/d/leg delivered in a concentration of 100p ul/ml</td>
</tr>
<tr>
<td></td>
<td>B.3 Out-of-study care</td>
<td>Patients will only be accepted by referral from physicians who will resume care of the patients after their participation in the protocol has concluded and who, if necessary, can assist in monitoring their warfarin therapy</td>
</tr>
<tr>
<td>B. Study Procedure</td>
<td>B.4 Measurement of drug level/pharmacodynamics/pharmacokinetics</td>
<td>Pharmacokinetic (PK) and pharmacodynamics (PD) sampling for dabigatran will occur on Days 0 – 1, Day 19±1 – 20, and Day 26±1 – 27. The PD effects of dabigatran will be characterized via ecarin clotting time measurements.</td>
</tr>
<tr>
<td></td>
<td>B.5 Provision of drug for study</td>
<td>The medications will be provided by the NIH or the drug manufacturer without cost to the patient for the duration of the study</td>
</tr>
<tr>
<td>B. Study Procedure</td>
<td>B.6 Use of other drugs</td>
<td>Supportive therapies, such as drugs for pain management, anti-emetics are to be administered when required by the patient’s condition.</td>
</tr>
<tr>
<td>--------------------</td>
<td>------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>B. Study Procedure</td>
<td>B.7 Assessment of drug effect</td>
<td>Irinotecan efficacy will be assessed by comparing drug and metabolite concentrations in tumor tissue.</td>
</tr>
<tr>
<td>C. Relation of Drug to Study</td>
<td>C.1 Study of drug effect</td>
<td>This study will test the effectiveness of low-dose recombinant tissue plasminogen activator in dissolving blood clots in deep leg veins.</td>
</tr>
<tr>
<td>C. Relation of Drug to Study</td>
<td>C.2 Relationship between study drug and safety/adverse event</td>
<td>Safety of therapy with Irinotecan-Drug Eluting Beads will be studied.</td>
</tr>
<tr>
<td>C. Relation of Drug to Study</td>
<td>C.3 Study of interaction between drugs</td>
<td>This investigation will use an open-label, fixed sequence design to evaluate the effect of P-gp inhibition by ritonavir or cobicistat on dabigatran etexilate pharmacokinetics and pharmacodynamics.</td>
</tr>
<tr>
<td>C. Relation of Drug to Study</td>
<td>C.4 Use of drug affects the study (for example, interacts with study drug) or interacts with the study (allowed, forbidden, reason for disqualification; this relates to events after enrollment)</td>
<td>If their current care involves botulinum toxin injections or taking baclofen which are listed as exclusions in the eligibility criteria, subjects cannot be in the study.</td>
</tr>
<tr>
<td>C. Relation of Drug to Study</td>
<td>C.5 Use of drug is related to subject enrollment (allowed, required, forbidden, reason for exclusion)</td>
<td>Patients on chronic nitrates, such as nitroglycerin, are excluded from the study.</td>
</tr>
<tr>
<td>C. Relation of Drug to Study</td>
<td>C.6 Patient must agree to some action related to taking drug</td>
<td>Women who are breast feeding must be willing and able to interrupt breastfeeding during the administration of filgrastim and for two days following the final dose.</td>
</tr>
<tr>
<td>C. Relation of Drug to Study</td>
<td>C.7 Patient allergy/intolerance to drug affects enrollment (allowed, required, reason for exclusion; this refers to the patient's response, not to a property of the drug)</td>
<td>Patients with severe allergic reactions to iodine contrast which cannot be controlled by premedication with antihistamines and steroids are not eligible.</td>
</tr>
<tr>
<td>C. Relation of Drug to Study</td>
<td>C.8 Rule regarding use of drug during study (allowed, not allowed; this relates to restrictions on the drug use, not changes to the patient status for using the drug)</td>
<td>Patients are prohibited from receiving other experimental agents/adjuvant treatments during the study.</td>
</tr>
</tbody>
</table>

Ontology Evaluation

The extraction of text samples from the test set produced 102 samples. Initial coding by both coders working independently resulted in 11 discrepancies (89% agreement), although VH suggested possible alternatives in two cases that matched JJC's classification (91% partial or complete agreement). After further discussion of the 18 concepts, the coders agreed on 101 cases (99%). The 11 discrepancies and their resolution are shown in Table 2.
Table 2. Coding discrepancies from test set (codes in parenthesis are initial alternate suggestions by Coder 2).

<table>
<thead>
<tr>
<th>Text Sample</th>
<th>Coder 1 (JJC)</th>
<th>Coder 2 (VH)</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>The NIH cannot guarantee that it[sic] will have continued access to the drug</td>
<td>B.5</td>
<td>B.2</td>
<td>B.5</td>
</tr>
<tr>
<td>Patients with hypokalemia despite medical therapy with mineralocorticoid antagonists will also be exclude from…the studies</td>
<td>No code</td>
<td>C.8</td>
<td>Coder 1 feels this is about a patient condition; Coder 2 feels this is about the medication use</td>
</tr>
<tr>
<td>Patients receive an intranasal application of OT, AVP or placebo and participate in a threat-of-shock experiment</td>
<td>B.2</td>
<td>C.1 (B.2)</td>
<td>B.2</td>
</tr>
<tr>
<td>Contraindications: Life-threatening reactions to previous influenza vaccination</td>
<td>C.7</td>
<td>C.5</td>
<td>C.7</td>
</tr>
<tr>
<td>Cannot be on any multivitamins or nutritional supplements</td>
<td>C.5</td>
<td>C.8</td>
<td>C.5</td>
</tr>
<tr>
<td>Is there a reasonable possibility the IP [investigational product] caused the event?</td>
<td>B.7</td>
<td>A.3 (B.7)</td>
<td>B.7</td>
</tr>
<tr>
<td>Exclusions: history of hypersensitivity reactions to cyclosdextrin</td>
<td>C.7</td>
<td>C.5</td>
<td>C.7</td>
</tr>
<tr>
<td>Prior human experience of intrathecal injection of HP-β-CD up to 350 mg suggest that the risk may be low.</td>
<td>A.3</td>
<td>A.2</td>
<td>A.3</td>
</tr>
<tr>
<td>All information concerning use of HP-β-CD in the context of this trial is considered confidential.</td>
<td>B.1</td>
<td>A.1</td>
<td>B.1</td>
</tr>
<tr>
<td>The main purpose of the study is to establish a safe dose of the drug…</td>
<td>C.2</td>
<td>C.1</td>
<td>C.2</td>
</tr>
<tr>
<td>Number of patients presenting with acute or delayed toxicities to each drug will be assessed during the whole treatment period…</td>
<td>C.2</td>
<td>C.1</td>
<td>C.2</td>
</tr>
</tbody>
</table>

Discussion

Contribution of the Study

This study identified three general and 18 specific roles that drugs may have in research studies. Not surprisingly, some of these roles are already represented in existing clinical research ontologies. In particular, ERGO covers the representation of drug use that are required for subject inclusion in a study or that result in subject exclusion from a study. However, additional relationships that involved drug administration or exclusion after subject enrollment.

While these ontologies can be expanded through theoretical mechanisms (for example, thought experiments about how drug terms might appear in study protocols), the empiric approach of studying actual protocols, taken by us and others,\(^3^,\(^4^)\) can at least serve as a reality check of an ontology’s completeness and may produce findings that might not have been anticipated through a purely theoretical approach. Our study contributes to this effort through review of a fairly substantial set of protocols.

Limitations

Despite the review of a large set of documents, a relatively small set of contexts was found. While reviewer fatigue may have been a contributing factor, the evaluation phase of our study found that an independently selected set of contextual statements was fully classifiable with the modest set discovered in the development phase of the study.

The development of the 18 roles was based on contextual phrases found in study protocols obtained exclusively from the NIH. The repetitive nature of the language found in the protocols, across 13 institutes and centers, dozens of specialties and hundreds (or perhaps thousands) of study authors suggests some possible cultural and habitual
effects. However, part of the test set (but not the training set) was drawn from protocols entered into the ClinicalTrials.gov database, which reflects a national pool of research. This suggests that the ontology thus developed is not NIH-specific.

Finally, this work does not directly provide an extension to any particular ontology and duplicates some previous work by others. Additional work is necessary to represent the knowledge gained in this study in a more formal manner. The ERGO framework\(^3\) appears to be a suitable path forward.

Potential Applications of Drug Context Information

Based on our own work at examining the ways in which drugs may adversely affect study procedures,\(^5\) it is clear that even the discovery of free-text phrases in relevant contexts can be useful, since they can serve as the basis for alerting messages to remind clinicians of study-specific requirements for patient management (e.g., “this drug should only be given to the subject to treat certain autoimmune disorders”). Deeper representation of such contexts (for example, allowing alert logic to handle exceptions to the alert) may be more elusive. However, even simple information can be actionable and given the richness of contexts of drugs in research studies, knowing when a drug that is mentioned in a protocol is *not* in a particular context may be valuable as well. It may therefore be useful to speculate on the kinds of capabilities that are possible by knowing more specific information about a drug’s context in a protocol. Table 3 provides some examples and the reader is encouraged to consider others.

Conclusion

We reviewed a large set of protocol documents in order to contribute to knowledge about the roles of drugs in research protocols. Our findings can serve as the basis for subsequent work on the development of clinical research ontologies for use in improving the quality of the research process through improved planning and execution of studies with, for example, study-relevant forms, orders sets and alerts.

Acknowledgments

This work was supported in part with intramural research funds from the NIH Clinical Center and the National Library of Medicine. The authors thanks Ida Sim, Samson Tu and Simona Carini for helpful information regarding OCRe and ERGO, in particular the sharing of the data set used for the ERGO study.

References

Table 3. Potential uses of knowledge about different roles of drugs in research studies.

<table>
<thead>
<tr>
<th>Role</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.3  Adverse event information</td>
<td>Improve thorough planning of monitoring and managing adverse events, based on information in same or similar protocols</td>
</tr>
<tr>
<td>B.1  Collection/retention of data about drug use</td>
<td>Improvement of case report forms, encounter forms, and clinical trials data management system templates</td>
</tr>
<tr>
<td>B.2  Administration and dosing of drug</td>
<td>Development of order sets for clinician order entry</td>
</tr>
<tr>
<td>B.3  Out-of-study care</td>
<td>Development of order sets for clinician order entry; letters to referring physicians</td>
</tr>
<tr>
<td>B.4  Measurement of drug level/pharmacodynamics/pharmacokinetics</td>
<td>Same as B.2</td>
</tr>
<tr>
<td>B.5  Provision of drug for study</td>
<td>Resource planning</td>
</tr>
<tr>
<td>B.6  Use of other drugs</td>
<td>Development of order sets for clinician order entry; development of alerts; resource planning</td>
</tr>
<tr>
<td>B.7  Assessment of drug effect</td>
<td>Same as B.2</td>
</tr>
<tr>
<td>C.1  Study of drug effect</td>
<td>Improving relevance of searches of studies that are related to a specific aspect of use of the drug for enrollment or secondary analysis of study results</td>
</tr>
<tr>
<td>C.2  Relationship between study drug and safety/adverse event</td>
<td>Same as A.3</td>
</tr>
<tr>
<td>C.3  Study of interaction between drugs</td>
<td>Development of decision support rules to prevent adverse events and protocol violations</td>
</tr>
<tr>
<td>C.4  Use of drug affects the study or interacts with the study</td>
<td>Same as C.3</td>
</tr>
<tr>
<td>C.5  Use of drug is related to subject enrollment</td>
<td>Establishing subject eligibility based on past drug administration data</td>
</tr>
<tr>
<td>C.6  Patient must agree to some action related to taking drug</td>
<td>Improvement of patient consent forms</td>
</tr>
<tr>
<td>C.7  Patient allergy/intolerance to drug affects enrollment</td>
<td>Establishing patient eligibility based on past intolerance data</td>
</tr>
<tr>
<td>C.8  Rule regarding use of drug during study</td>
<td>Same as C.3; also, development of order sets for clinician order entry</td>
</tr>
</tbody>
</table>
Improving EHR Capabilities to Facilitate Stage 3 Meaningful Use Care Coordination Criteria

Dori A. Cross, BSPH1, Genna R. Cohen, BS1, Paige Nong, BS1, Anya-Victoria Day, MPH2, Danielle Vibbert, MPH2, Ramya Naraharisetti, M.Sc.2, Julia Adler-Milstein, PhD1,3

1 School of Public Health, University of Michigan, Ann Arbor, MI; 2 Altarum Institute, Ann Arbor, MI; 3 School of Information, University of Michigan, Ann Arbor, MI

Abstract

Primary care practices have been limited in their ability to leverage electronic health records (EHRs) and health information exchange (HIE) to improve care coordination, but will soon be incentivized to do so under proposed Stage 3 meaningful use criteria. We use mixed methods to understand how primary care practices manage, share and reconcile electronic patient information across care settings, and identify innovations in EHR design to support enhanced care coordination. Opportunities identified by practices focused on availability and usability of features that facilitate (1) generation of customized summary of care records, (2) team-based care approaches, and (3) management of the increased volume of electronic information generated and exchanged during care transitions. More broadly, vendors and policymakers need to continue to work together to improve interoperability as the key to effective care coordination. If these EHR innovations were widespread, the value of meeting the proposed Stage 3 care coordination criteria would be substantially enhanced.

Introduction

Background: The centerpiece of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 is financial incentives for providers who implement electronic health records (EHRs) and use them in accordance with federally-specified meaningful use criteria.1 Criteria that promote care coordination are heavily emphasized because of the potential for EHRs, coupled with electronic health information exchange (HIE), to enable sharing of patients’ health information between delivery settings and ultimately improve the quality, safety, and efficiency of care. These criteria were, however, largely deferred to later stages of meaningful use because few providers had EHRs that were capable of exchanging or using electronically shared clinical data for effective care coordination.2,3,4

Objectives: We sought to identify EHR capabilities and innovations that would better enable primary care practices to meet the proposed Stage 3 care coordination criteria in a meaningful way.

Methods

Our study focused on the proposed Stage 3 meaningful use care coordination criteria from the perspective of primary care practices. At the time the study was designed, there were three proposed care coordination criteria:

1. SGRP 302: Reconcile patients’ medications for more than 50% of transitions of care and reconcile patients’ medication allergies and problems (e.g., uncontrolled diabetes) for more than 10% of transitions in care.

2. SGRP 303: Send a summary of care record (SCR) for at least 65% of transitions of care or referrals with at least 30% sent electronically. [SCRs for referrals must include a “concise narrative in support of care transitions,” i.e., free text that captures the current care synopsis and expectations for referral.]

3. SGRP 305: Be prepared to receive an EHR-generated acknowledgement from practices that received information from the focal practice for at least 50% of referrals, i.e., a referral receipt, with at least 10% returned electronically.

Design and Setting

We use mixed quantitative and qualitative methods to identify technical barriers faced by primary care practices in the state of Michigan that are pursuing proposed Stage 3 meaningful use care coordination criteria, and develop recommendations for EHR innovations that would best enable practices to meet the criteria in ways that improve care coordination. Our study included a statewide survey of 328 primary care practices, complemented by three
rounds of interviews in 12 representative primary care practices (the implementation sample) working towards proposed Stage 3 care coordination criteria. Sample descriptive characteristics are reported in Table 1. All practices in both samples had achieved Stage 1 Meaningful Use (as of September 1, 2013) with support of the Michigan Center for Effective IT Adoption (M-CEITA), the Michigan Regional Extension Center.

Table 1. Characteristics of Primary Care Practices in Survey and Implementation Samples

<table>
<thead>
<tr>
<th>Practice Size</th>
<th>Statewide Survey Sample (n=328 practices)</th>
<th>Implementation Sample (n=12 practices)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small (2 or fewer)</td>
<td>45%</td>
<td>58%</td>
</tr>
<tr>
<td>Medium (2-5)</td>
<td>36%</td>
<td>17%</td>
</tr>
<tr>
<td>Large (6 or more)</td>
<td>19%</td>
<td>25%</td>
</tr>
<tr>
<td>Practice is Independently Owned</td>
<td>56%</td>
<td>92%</td>
</tr>
<tr>
<td>Practice is Affiliated with a Physician Organization*</td>
<td>88%</td>
<td>42%</td>
</tr>
<tr>
<td>Duration of EHR Use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2 years</td>
<td>32%</td>
<td>33%</td>
</tr>
<tr>
<td>3-5 years</td>
<td>38%</td>
<td>17%</td>
</tr>
<tr>
<td>6+ years</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>Number of Different EHR Vendors Represented in the Sample</td>
<td>41</td>
<td>9</td>
</tr>
</tbody>
</table>

* In Michigan physician organizations are umbrella organizations—such as independent practice organizations, physician-hospital organizations, and large multispecialty group practices—that provide clinical leadership, administrative structure, technical infrastructure, and other resources for physician practices.

Data Collection

Survey: The statewide survey captured practice demographics, readiness for Stage 2 and Stage 3 meaningful use care coordination criteria, health information exchange (HIE) participation, facilitators and barriers to meeting Stage 3 criteria, perceived impact of Stage 3 criteria and optimal approach to information sharing to support care coordination. Some questions were targeted to the practice manager (PM) and others were targeted to a primary care provider (PCP) in the practice.

A random sample of 328 primary care practices completed the survey between November 2013 and March 2014. Participants were offered multiple means to complete the survey: phone, online (Qualtrics), or on paper via fax or mail. We received responses from 233 practice managers (71% response rate) and 174 primary care providers (53% response rate). Data were imported and analyzed in STATA 12.0. We estimated all figures using survey sampling weights based on our sampling strategy in order to generalize results to the statewide population of primary care practices that had achieved Stage 1 meaningful use.

Implementation Sample: We selected 12 practices to receive technical assistance to facilitate achievement of proposed Stage 3 meaningful use care coordination criteria. We conducted three rounds of semi-structured interviews with key practice staff (the practice manager and at least one PCP) between October 2013 and June 2014; interviews were in person at the outset of implementation (i.e., before attempting to achieve the Stage 3 criteria), by phone three months later, and again in person six months following initiation of implementation. The initial round of interviews focused on current state processes of supporting care coordination using EHRs. The second round of interviews focused on barriers to achieving Stage 3 care coordination measures and potential strategies to overcome them. The final round of interviews focused on progress towards achieving the criteria, suggested changes to the criteria, strategies for increasing the impact of the criteria, and EHR innovations to support criteria achievement. Interview guides were developed using the Consolidated Framework for Implementation Research, which provides a pragmatic structure for organizing key domains across published implementation theories (1, 2). It is particularly well suited to technology evaluation because it addresses the complex, interacting, multi-level, and transient nature
of phenomenon in real-world healthcare settings. Teams of two researchers trained and experienced in qualitative interviewing conducted all interviews.

**Data Analysis**

From the survey data, we generated descriptive statistics to capture current EHR and HIE capabilities related to care coordination as well as the ability to meet proposed Stage 3 coordination criteria. From the interview data, we sought to understand how EHRs and other types of health IT were used to manage patient care across delivery settings, and identify opportunities for innovation in EHR design to support enhanced care coordination. Codes were developed a priori based on constructs hypothesized to be relevant based on the literature and were expanded as needed to incorporate emergent themes. Two researchers independently coded all transcripts and met to reconcile codes through discussion. Final coded transcripts were coded imported into Atlas.ti and analyzed to extract key findings after each phase of interviewing.

**Results**

*EHR Innovation to Improve Summary of Care Records (SCRs)*

Practices in the implementation sample had EHRs capable of generating Summary of Care Records, and most utilized the local health information exchange effort (Michigan Health Connect) to send them electronically during a referral by uploading the SCR to a portal. A common challenge was that SCRs contained a lot of superfluous information that was auto-generated by the EHR and this interfered with clinicians’ ability to locate relevant information. Primary care providers (PCPs) felt that specialists often missed important details relevant for the referral because they were so difficult to find in the SCR. As a result, one-third of practices inconsistently or never used the auto-generate SCR feature within their EHR.

Data from our statewide survey revealed varied opinions across PCPs as to which record elements should be shared when patients are transferred across care settings (Table 2). Practices in the implementation sample further explained that decisions about the relevancy of certain information also varied within provider, based on the patient, his/her history and preferences of the receiving physician. While federal criteria will dictate the types of information that must be shared to support transitions of care, our data point to the value of designing EHR functionality that enables more customization of SCRs.
Table 2. Percent of Primary Care Providers Responding that Specific Information Elements should be Shared during Transitions of Care (N=174 PCPs)

<table>
<thead>
<tr>
<th>Information Element</th>
<th>REFER a patient to a specialist</th>
<th>RECEIVE a patient back from a specialist</th>
<th>RECEIVE a patient after discharge from the hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem list</td>
<td>77%</td>
<td>59%</td>
<td>73%</td>
</tr>
<tr>
<td>Assessment (e.g., notes summarizing key problems)</td>
<td>76%</td>
<td>78%</td>
<td>75%</td>
</tr>
<tr>
<td>Medication allergies</td>
<td>75%</td>
<td>60%</td>
<td>60%</td>
</tr>
<tr>
<td>Radiology reports</td>
<td>75%</td>
<td>76%</td>
<td>76%</td>
</tr>
<tr>
<td>Lab test results</td>
<td>74%</td>
<td>78%</td>
<td>78%</td>
</tr>
<tr>
<td>Known contra-indications for medications patient is taking</td>
<td>64%</td>
<td>55%</td>
<td>59%</td>
</tr>
<tr>
<td>Care plan</td>
<td>44%</td>
<td>79%</td>
<td>75%</td>
</tr>
<tr>
<td>Radiology images</td>
<td>40%</td>
<td>39%</td>
<td>36%</td>
</tr>
<tr>
<td>Social history</td>
<td>35%</td>
<td>16%</td>
<td>16%</td>
</tr>
<tr>
<td>Assessment of functional status (e.g., ability to perform ADLs)</td>
<td>35%</td>
<td>45%</td>
<td>56%</td>
</tr>
</tbody>
</table>

Approximately one-quarter of implementation sample practices reported that their EHR currently included some functionality to customize SCRs, such as the ability to limit information by date range, check-boxes or radio buttons to deselect certain types of information, and/or templates to consistently pull the same information for specific specialists or referral types. A few providers also designed customizations with their vendor to have flexibility in structuring the document so that they could prioritize a concise narrative and emphasize the most relevant details. These features typically included moving the narratives to the front of the document and other visual cues (e.g. highlighting or bolding) to indicate importance.

Practices in the implementation sample felt that SCR functionalities could be further improved by enhancing usability and degree of available customization. For example, practices sought better filtering options to specify a more granular level of patient history to be included in the SCR while still meeting meaningful use criteria. Practices also felt that branching logic or guided workflows would make it easier to streamline the creation of customized SCRs.

Primary care providers valued receiving SCRs from specialists. However, they cited similar challenges in that the SCRs generated by specialists’ EHRs were often dense and obscured relevant new information. PCPs felt that SCRs that visually highlighted new or updated information generated during the referral would help them review the SCR and identify what to incorporate into their EHR.

**Current SCR Features Valued for Care Coordination:**
- Customization/electronic editing of information contained in SCRs through use of check-boxes, date range filters and templates
- Strategic location of concise narrative; reordering of SCR components to emphasize important information

**Future SCR Innovations to Enhance Care Coordination:**
- Ability to better filter information for a more concise and usable SCR, while still meeting meaningful use criteria
- Branching logic or guided workflows to facilitate customization of SCRs
- SCR generation from primary care to specialists: Functionality to visually or otherwise highlight most important pieces of information throughout the document
EHR Innovation to Support Team-based Primary Care Delivery

A broader set of policy efforts beyond meaningful use promotes enhanced primary care delivery models that have significantly increased care coordination and documentation requirements for primary care practices. In response, many implementation sample practices described transitioning to team-based care models that better utilize staff at different levels of training in order to meet patient care needs without additional physician burden. Implementation sample practices described a need for their EHR to better accommodate and support the documentation, workflows and necessary communication tools of multi-disciplinary care teams. Specifically, EHR innovations are needed that improve work coordination among staff and increase transparency of accountability when multiple individuals are providing care to the same patient.

The most common EHR feature used to support team-based care was tasking and task management. Over two-thirds of practices used tasking to track patient care needs as well as to delegate and follow up on staff responsibilities. Some EHR task features were more robust, allowing providers and staff to indicate status of a task, and run reports or receive alerts when tasks were incomplete. Many practices coupled this functionality with use of internal notes or messaging functions to create a record of task hand-offs and provide a communication space for shared understanding of responsibility. Having a team operate off of an assigned task queue served to coordinate tasks and also kept workloads open and transparent, providing accountability and ensuring tasks were completed in a timely way.

A small subset of implementation sample practices used embedded templates and clinical decision support features to facilitate greater autonomy and scope of work for non-physician staff. For example, templates and guided workflows with pop-up decision support helped one practice better utilize their nurses and mid-level staff in more enhanced care roles. Additional features, such as the ability to rank tasks by sensitivity or deadline, were requested by several practices. Practices also called for further EHR innovations to accommodate and integrate the workflow and documentation needs of ancillary team members, such as care managers, who are increasingly being incorporated into care teams.

Current Team-based Care Features Valued for Care Coordination:
- Robust tasking features with shared queues, direct assignment and alerts to identify incomplete tasks
- Plural patient record access to support concurrent workflows of team members
- Ability to electronically transmit notes to other team members within the EHR for improved communication
- Templates and guided workflows with clinical decision support to enable more delegation and greater autonomy to lower-level staff

Future Team-based Care Innovations to Enhance Care Coordination:
- Enhanced task functionality such as on-screen to-do lists and ability to rank tasks by sensitivity or deadline
- Functionalities to accommodate and integrate workflow and documentation needs of ancillary team members, such as care managers

EHR Innovation to Better Manage Patient Care in an Information-Rich Context

Richer information flowing between and within practices has the potential to enhance primary care provider decision-making and improve care coordination. In the absence of interoperability across EHR vendor systems, however, practices struggled to effectively manage this information and therefore felt they were not fully capturing the potential benefits of their EHR. EHR innovations to better manage patient care in an information-rich context fell into two areas: receipt of information and reconciliation of information.

EHR functionality for receiving information during patient care transitions

Referrals: All practices hoped that interoperability across EHR vendor systems would enable automated updating of information within their EHR. Specific to referrals, practices wanted their EHR to contain current referral status and
information from referral reports. Practices that were part of an integrated health system and on a shared EHR platform had these features and described more timely notification about referral status, and better integration of information from those encounters into their EHR.

Despite a lack of interoperability, some practices’ EHRs did a better job of supporting referral workflow and streamlining staff processes to better manage patients’ care transitions. For example, some practices used task lists or orders to track referrals or manage communication with specialists. These items were kept “open” in the EHR until the referral report was received. Staff found it particularly useful to be able to run reports on pending or overdue communication from the specialist regarding appointments or visit notes. Maintaining an accurate status of referrals, however, required substantial human-mediated communication with specialists.

Practices also requested better tools to incorporate information from external settings into their EHR in a consistent and easily accessible way. Providers were often unaware that new information was available because the EHR offered multiple options for where the information might be stored and it was not obvious when new information had been entered. Providers therefore failed to use documentation sent from external settings to inform care. Design elements such as the ability to tag or label documents for cross-listing in multiple places within the EHR would help providers more easily locate and utilize information.

Hospitalizations: Many practices relied on hospital portals to learn if their patients were hospitalized and to download documentation about their patients. Some practices received automated admission, discharge, or transfer (ADT) alerts, but felt overwhelmed by the volume of separate notices received throughout a patient’s hospital visit and post-hospitalization. Providers and staff were most interested in notice of admission and details contained in the discharge summary; intermediate notifications were considered less useful. Threading of hospital communications and ability to have EHR systems recognize and flag incoming communication by level of importance were therefore perceived as valuable.

**Current Data Management Features Valued for Care Coordination:**

- Ability for practices to run reports on pending or overdue tasks/orders to manage and close the loop on patient referrals

**Future Data Management Innovations to Enhance Care Coordination:**

- Automatic incorporation of referral reports, other incoming information, into relevant section of patient record
- Document tagging and/or ability to cross-list documents in multiple EHR locations for easier information retrieval
- More automation in closing out pending tasks/orders for completed referrals, e.g. when referral reports are scanned in
- Enhanced notification system for arrival of new information in the EHR, including ability to filter/sort by importance

**EHR functionality for managing information following patient care transitions**

Reconciliation: Practices struggled to incorporate large volumes of incoming documentation into their EHR. Practices expressed desire for auto-reconciliation features, with some automatic import and integration of data from hospitals and specialists. With the exception of labs and imaging, practices were still largely receiving read-only files – via fax, mail, or hospital portal download – that require scanning into the EHR. Even documents received electronically through the community HIE platform or via e-fax were still read-only documents; reconciliation therefore required a visual comparison of records with manual entry of new information.

A frustrating result of this process was multiple entries of the same (or similar) diagnoses or prescriptions that were entered into the EHR, cluttering the problem and medication lists. EHR auto-reconciliation functionality – the ability to support de-duplication or collapsing of similar entries to produce cleaner, more usable lists – was therefore perceived as valuable. Practices felt that the ability to easily tie ICD codes to entries within both the problem and medication lists would be one useful way to provide underlying structure for this functionality. Linking entries to the ICD coding classification system would also enable better organization of existing information within the record and provide the basis for enhanced interaction and safety alerts.
Some practices had the option to embed links to ICD codes using structured fields; however, practices reported inconsistent use of the functionality. Clinicians struggled to search underlying databases to locate and attach the appropriate ICD code; currently available software did not provide intuitive, comprehensive or consistently accurate ICD search capability. While auto-reconciliation of internal records with documentation from specialists, hospitals or third-party sources of pharmacy information is not possible in the absence of interoperability, practices felt that receiving incoming documentation with attached ICD codes would still provide value by aiding their staff in performing manual reconciliation.

**Acute vs. Chronic Documentation:** At least half of implementation sample practices described challenges related to accurate and efficient documentation of short-term, acute problems or medications in their EHRs. Providers sought EHR enhancements that would improve their ability to capture the distinction between active versus past medications, and acute versus chronic conditions when managing and sending out patient information. Lack of this functionality resulted in providers making documentation decisions likely to compromise patient care. For example, some practices chose not to enter acute problems, such as sinus infections, in the problem list to avoid clutter. This information may be valuable to specialists and without it providers may fail to identify chronic issues.

Several EHRs allowed providers to enter course of treatment information for prescriptions so that medications would automatically drop off of the “current” medication list. Providers valued drag-and-drop functionality and radio buttons to easily move entries between the active and past sections of problem and medication lists.

**Current Documentation Features Valued for Care Coordination:**
- Auto-removal of short-term medications from active medication list based on entered course of treatment
- Drag-and-drop or radio button features to easily move problems or medications to past information section(s)

**Future Documentation Innovations to Enhance Care Coordination:**
- Ability to attach ICD codes to problems and medications, enabling:
  - easier reconciliation
  - threading/grouping of similar issues
  - improved interaction alerts and decision support
- Tailored functionality for differentiating acute versus chronic problems and medications

**Discussion**

Despite significant increases in EHR adoption since the passage of HITECH, most primary care providers have little experience exchanging or using electronically shared clinical data to support care coordination and associated improvements in care quality, safety and efficiency. This study is one of the first to examine practices’ capabilities and use of EHRs and HIE to support care coordination in the context of proposed Stage 3 meaningful use objectives. Findings reveal key ways in which EHR and HIE functionality is already being used to support care coordination, and areas of opportunity for further innovation. Specifically, primary care practices called for improved availability and usability of features that facilitate (1) generation of customized SCRs, (2) team-based care approaches, and (3) management of the increased volume of electronic information generated and exchanged during patient care transitions.

Some of these innovations are likely much easier than others for EHR vendors to implement; for example, improved filtering mechanisms or highlighting functionality to improve SCR customization. Other innovations such as incorporating ICD codes into problem and medication lists may be more difficult to implement, or, in the case of widespread interoperability, are difficult for EHR vendors to tackle individually. For these, policymakers should consider including them in certification criteria, and, in parallel, consider creating communities of practice that enable EHR vendors to come together (perhaps with providers as well) and work to refine and specify the innovations. Policymakers and vendors must also remain cognizant of not just availability but also usability of different EHR features. Increased focus by developers on human-centered design methods and use of workflow analysis to inform the development of these enhanced EHR functionalities will be critical to realizing intended benefits of IT adoption.
Bolstered by Stage 3 meaningful use criteria and simultaneous delivery system reform efforts such as patient-centered medical homes and accountable care organizations, there will be increased pressure on EHRs to support improved care coordination. EHR vendors, guided by market demand and policy efforts, play a critical role in ensuring that practices have the technical capability – and usability – to enhance their care management practices. Beyond adding or enhancing system features, vendors are a key stakeholder in national conversations around achieving greater interoperability. Going forward, vendors and policymakers need to work in concert to create conditions in which all stakeholders see the benefit of and are willing to invest in robust information exchange capability in order to achieve optimal care coordination.

**Conclusion**

The ability of practices to effectively use EHRs and electronic information exchange to enhance care coordination practices is essential to translate the large national investment in health IT into improved care and patient outcomes. This study examines practices’ current capabilities and use of EHRs to support care coordination in the context of Stage 3 meaningful use objectives, with the specific goal of providing key recommendations for how to enhance EHR functionality. Vendors ultimately need to work with policymakers to find a viable approach to interoperability and create market conditions in which secure electronic sharing of patient information can be achieved to improve care coordination.

**References**

Abstract

Biomedical ontologies play a vital role in healthcare information management, data integration, and decision support. Ontology quality assurance (OQA) is an indispensable part of the ontology engineering cycle. Most existing OQA methods are based on the knowledge provided within the targeted ontology. This paper proposes a novel cross-ontology analysis method, Cross-Ontology Hierarchical Relation Examination (COHeRE), to detect inconsistencies and possible errors in hierarchical relations across multiple ontologies. COHeRE leverages the Unified Medical Language System (UMLS) knowledge source and the MapReduce cloud computing technique for systematic, large-scale ontology quality assurance work. COHeRE consists of three main steps with the UMLS concepts and relations as the input. First, the relations claimed in source vocabularies are filtered and aggregated for each pair of concepts. Second, inconsistent relations are detected if a concept pair is related by different types of relations in different source vocabularies. Finally, the uncovered inconsistent relations are voted according to their number of occurrences across different source vocabularies. The voting result together with the inconsistent relations serve as the output of COHeRE for possible ontological change. The highest votes provide initial suggestion on how such inconsistencies might be fixed. In UMLS, 138,987 concept pairs were found to have inconsistent relationships across multiple source vocabularies. 40 inconsistent concept pairs involving hierarchical relationships were randomly selected and manually reviewed by a human expert. 95.8% of the inconsistent relations involved in these concept pairs indeed exist in their source vocabularies rather than being introduced by mistake in the UMLS integration process. 73.7% of the concept pairs with suggested relationship were agreed by the human expert. The effectiveness of COHeRE indicates that UMLS provides a promising environment to enhance qualities of biomedical ontologies by performing cross-ontology examination.

Introduction

An ontology provides a formal description of entities (or concepts) and relationships between entities in a given domain. Ontologies or controlled terminologies in biomedical domains, such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9CM), play a vital role in healthcare information management. For example, SNOMED CT has been designated in the U.S. and many other countries as the preferred terminology for use of electronic health records (EHRs) [1]. Biomedical ontologies also serve as essential knowledge sources for many biomedical applications including text processing [2, 3], data integration [4, 5], decision support [6], reasoning [7], and information retrieval [8].

Most biomedical ontologies have a large size and a complex structure. It is unavoidable that errors and inconsistencies may be introduced. Hence researchers have proposed auditing methods to detect possible errors and inconsistencies in concepts or relationships to enhance the quality of biomedical ontologies. Jiang and Chute [9] leveraged Formal Concept Analysis to identify missing concepts in SNOMED CT. Zhang and Bodenreider [10, 11] proposed a lattice-based approach for exhaustive auditing of SNOMED CT structure. Gu et al. [12] presented computable methods to detect erroneous relationships in the Foundational Model of Anatomy (FMA). Luo et al. [13] developed an automated method to audit symmetric concepts in FMA using self-bisimilarity and linguistic features of the concept labels. Ochs et al. [14] presented a subject-based abstraction network method to audit SNOMED CT using the relationships provided. Mortensen et al. [15] evaluated crowdsourcing methods to detect errors in SNOMED CT. Zhang et al. [16, 17] developed fast MapReduce algorithms for lattice-based evaluation with dramatically reduced time. We refer to [18] for a more detailed review of early auditing methods for biomedical terminologies.

Most existing works have focused on developing auditing methods utilizing the knowledge (e.g., terms, relationships) provided within a targeted ontology. This paper proposes a novel cross-ontology method, Cross-Ontology Hierarchical Relation Examination (COHeRE), to detect possible errors and inconsistencies in hierarchical relationships across multiple biomedical ontologies in the Unified Medical Language System (UMLS) [19]. Inconsistencies uncovered using COHeRE involve UMLS concept pairs related by different types of relationships in disparate source vocabularies. For example, Prolonged depressive adjustment reaction is related to Adjustment disorder with depressed mood.
through the *is-a* relationship in SNOMED CT, while *Prolonged depressive reaction* is related to *Adjustment disorder with depressed mood* through the *sibling* relationship in ICD9CM (Figure 1). Since not all such relations can be correct, the inconsistencies provide immediate candidates for auditing and change. For example, the *sibling* relationship in Figure 1 should probably be changed to *is-a*, in ICD9CM.

![Figure 1: Example of inconsistent relations in SNOMED CT and ICD9CM integrated in the UMLS.](image)

COHeRE takes the UMLS concepts and relations as input, and produces inconsistent relations (and the number of distinct occurrences) as output. Within UMLS, 138,987 concept pairs were found to have inconsistent relationship types across multiple source vocabularies. 22,644 concept pairs were found to be related by *CHD* in some source vocabularies, but were specified as other relations (*RO*) in other vocabularies. 2,755 concept pairs were found to have the hierarchical child relation (*CHD*) in some source vocabularies, but being in a sibling relation (*SIB*) in others. Manual evaluation of 40 inconsistent concept pairs detected by COHeRE shows that 95.8% of the involved inconsistent relations indeed exist in their source vocabularies, and 73.7% of the concept pairs with suggested relationship by COHeRE are valid.

1 Background

1.1 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) [19], developed by National Library of Medicine (NLM), is a large integrated repository of biomedical source vocabularies to facilitate interoperability among disparate systems in medicine. Knowledge in the UMLS is organized by concept (or meaning). Term variants from source vocabularies are clustered together to form a concept, and each concept is assigned a unique concept identifier (CUI). For example, terms *Myocardial infarction*, *Infarction of heart*, *Heart attack*, *MI* and *Cardiovascular Stroke* from different sources represent the same meaning and are assigned a unique CUI: C0027051. The 2014AB release of the UMLS contains over 3 million concepts and 11.9 million unique concept names from more than 150 source vocabularies, including SNOMED CT, FMA, and ICD9CM. The concepts are provided in the distribution file MRCONSO of the UMLS.

Moreover, relationships between terms in source vocabularies are preserved in the UMLS as relationship attributes (RELA). There are over 600 relationship attributes in the UMLS, abstracted into 10 broader relationship types (REL). REL can be further classified as synonymous, hierarchical, or associative. For instance, *Coccyx Injury* (C0560630) and *Back Injury* (C0004601) are related through *is a* (RELA) in SNOMED CT. It is classified as a hierarchical relationship *CHD* (REL) in the UMLS, where *CHD* means child relationship. Table 1 shows the UMLS relationships (2014AB release) at different granularity. The 2014AB version of the UMLS contains over 60 million relations between terms, which are provided in the distribution file MRREL. In this paper, relationship refers to an individual link between two terms, while relationship indicates a type of relation.

1.2 MapReduce

We use a cloud-computing technique called MapReduce [20] in order to process large amounts of data required by COHeRE. MapReduce is a scalable distributed computing framework to process large amounts of data. A MapReduce program consists of a mapper and a reducer, to transform data in the form of key-value pairs. The mapper has a map method to transform input (*key, value*) pairs into intermediate (*key', value'*) pairs. The reducer has a reduce method to transform intermediate (*key', [value]*) aggregates into output (*key'', value'*) pairs. MapReduce has been successfully used for developing scalable approaches to auditing *is a* relationship in SNOMED CT [16, 17, 21, 22].
### Classes

<table>
<thead>
<tr>
<th>Class</th>
<th>REL</th>
<th>RELA Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>SY (source asserted synonymy)</td>
<td>same as, alias of</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>CHD (has child relationship in a source vocabulary)</td>
<td>is a, part of</td>
</tr>
<tr>
<td></td>
<td>PAR (has parent relationship in a source vocabulary)</td>
<td>inverse is a, has part</td>
</tr>
<tr>
<td></td>
<td>SIB (has sibling relationship in a source vocabulary)</td>
<td>sibling in is a, sibling in part of</td>
</tr>
<tr>
<td></td>
<td>RN (has a narrower relationship)</td>
<td>tradename of, form of</td>
</tr>
<tr>
<td></td>
<td>RB (has a broader relationship)</td>
<td>has tradesman, has form</td>
</tr>
<tr>
<td>Associative</td>
<td>AQ (Allowed qualifier)</td>
<td>actual outcome of, modifies</td>
</tr>
<tr>
<td></td>
<td>QB (can be qualified by)</td>
<td>has actual outcome, modified by</td>
</tr>
<tr>
<td></td>
<td>RQ (related and possibly synonymous)</td>
<td>classified as, classifies</td>
</tr>
<tr>
<td></td>
<td>RO (has relationship other than synonymous, narrower, or broader)</td>
<td>measured by, measures</td>
</tr>
</tbody>
</table>

Table 1: Relationships in the UMLS 2014AB release (REL: relationship, RELA: relationship attribute).

#### 2 Methods

Figure 2 shows the overview of the proposed pipeline of steps for COHeRE. First, the UMLS concepts and relations are taken as the input; relations are filtered according to certain criteria; and relations claimed in source vocabularies are aggregated for each concept pair. Second, inconsistent relations are detected if a concept pair has disparate relations in different source vocabularies. Finally, the inconsistent relations are voted by source vocabularies, and the voting result together with the inconsistent relations are the output.

![Figure 2: Overview of the proposed method COHeRE.](image)

#### 2.1 Relation Filtering and Aggregation

**Material preparation.** The MRCONSO (1.41GB) and MRREL (5.48GB) distribution files in the UMLS 2014AB release are used as the input for concepts and relations between concepts, respectively. Since UMLS is multilingual and covers a wide range of vocabularies, a subset of most relevant source vocabularies (Table 2) are manually selected to perform the cross-ontology examination for this study. Hence terms that are not in the subset of source vocabularies are discarded from MRCONSO. In addition, concepts with terms containing word *unspecified* or *NOS* (Not Otherwise Specified) are filtered out from MRCONSO, since such terms are very likely to cause error-prone relations in the integrated MRREL.

**Relation filtering.** Each relation record in MRREL can be represented as \((C_1, C_2, R, RA, S)\), where \(C_1\) and \(C_2\) form a concept pair, \(R\) is the relationship between \(C_1\) and \(C_2\) indicating \(C_2 \text{ R } C_1\), \(RA\) is the relationship attribute of \(R\), and \(S\) is the source vocabulary. Since every source relationship is represented in two directions in separate rows in UMLS, analyzing both directions of a concept pair is redundant. For example, if a concept pair \((C_1, C_2)\) is related through the \(CHD\) relationship with the *is a* relationship attribute, then \((C_2, C_1)\) is symmetrically related through the \(PAR\) relationship with the *inverse is a* relation attribute. To avoid duplicated analysis, only one direction \(CHD\) is used for analyzing this concept pair. Generally, concept pairs with relationship \(CHD, RN, SIB, SY, AQ\) or \(RO\) are used for the cross-ontology analysis. In addition, a concept pair is discarded if the concepts share the same unique identifier \(CUI\), since COHeRE will only investigate relationships between distinct concepts. Concept pairs with relationship attribute \(RA\) as *mapped from* or *mapped to* are also ignored, since they are generated by UMLS and represent mappings between source vocabularies. Due to the large size of the MREL file, this filtering process is implemented using MapReduce (Mapper in Figure 3) to enable parallel computing.

**Relation aggregation.** For each concept pair \((C_1, C_2)\), source relationships are aggregated into the format of

\[
(C_1, C_2) \rightarrow (R_1, RA_1, S_1) \mid (R_2, RA_2, S_2) \mid \ldots \mid (R_n, RA_n, S_n).
\]
Table 2: A list of relevant source vocabularies for cross-ontology examination.

<table>
<thead>
<tr>
<th>Vocabulary Acronym</th>
<th>Vocabulary Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATC</td>
<td>Anatomical Therapeutic Chemical Classification System</td>
</tr>
<tr>
<td>CDT</td>
<td>Code on Dental Procedures and Nomenclature</td>
</tr>
<tr>
<td>CPM</td>
<td>Columbia Presbyterian Medical Center Medical Entities Dictionary</td>
</tr>
<tr>
<td>CPT</td>
<td>Current Procedural Terminology</td>
</tr>
<tr>
<td>FMA</td>
<td>Foundational Model of Anatomy Ontology</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>HL7V2.5</td>
<td>Health Level Seven Vocabulary Version 2.5</td>
</tr>
<tr>
<td>HL7V3.0</td>
<td>Health Level Seven Vocabulary Version 3.0</td>
</tr>
<tr>
<td>ICD10</td>
<td>International Statistical Classification of Diseases and Related Health Problems</td>
</tr>
<tr>
<td>ICD10CM</td>
<td>International Classification of Diseases, 10th Edition, Clinical Modification</td>
</tr>
<tr>
<td>ICD9CM</td>
<td>International Classification of Diseases, Ninth Revision, Clinical Modification</td>
</tr>
<tr>
<td>ICNP</td>
<td>International Classification for Nursing Practice</td>
</tr>
<tr>
<td>MDR</td>
<td>Medical Dictionary for Regulatory Activities Terminology (MedDRA)</td>
</tr>
<tr>
<td>MEDCIN</td>
<td>Medcomp Systems</td>
</tr>
<tr>
<td>MEDLINEPLUS</td>
<td>MedlinePlus Health Topics</td>
</tr>
<tr>
<td>MSH</td>
<td>Medical Subject Headings</td>
</tr>
<tr>
<td>MTH</td>
<td>UMLS Metathesaurus</td>
</tr>
<tr>
<td>NCBI</td>
<td>NCBI Thesaurus</td>
</tr>
<tr>
<td>NCNP</td>
<td>National Drug File - Reference Terminology</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>RXNORM</td>
<td>RxNorm Vocabulary</td>
</tr>
<tr>
<td>SNOMEDCT_US</td>
<td>US Edition of SNOMED CT</td>
</tr>
<tr>
<td>UMD</td>
<td>Universal Medical Device Nomenclature System</td>
</tr>
<tr>
<td>VANDF</td>
<td>Veterans Health Administration National Drug File</td>
</tr>
</tbody>
</table>

---

1: **Input:** Relations in MRREL \( \{(C_1, C_2, R, RA, S)\} \)
2: **Output:** Aggregated relations by concept pairs:

\[
\{(C_1, C_2) \rightarrow (R_1, RA_1, S_1) | (R_2, RA_2, S_2) | \ldots | (R_n, RA_n, S_n)\}
\]

3: **class** Mapper
4: Setup a HashSet \( H \) and load it with all the relevant concepts in MRCONSO using DistributedCache.
5: **method** Map((\( C_1, C_2, R, RA, S \)))
6: if \( H \) contains both \( C_1 \) and \( C_2 \), \( C_1 \) and \( C_2 \) are not identical, \( R \) is not in \( \{CHD, RN, SIB, SY, AQ, RO\} \), and \( RA \) is mapped from or mapped to then
7: Emit((\( C_1, C_2 \)), (\( R, RA, S \)))
8: end if

9: **class** Reducer
10: **method** Reduce((\( C_1, C_2 \)), ((\( R_1, RA_1, S_1 \)), (\( R_2, RA_2, S_2 \)), \ldots , (\( R_n, RA_n, S_n \))))
11: Setup a HashMap \( M \) for source vocabularies and relationships.
12: if \( C_1 \) and \( C_2 \) are not multiply related for each source vocabulary in \( M \) then
13: Emit((\( C_1, C_2 \)), ((\( R_1, RA_1, S_1 \)), (\( R_2, RA_2, S_2 \)), \ldots , (\( R_n, RA_n, S_n \))))
14: end if

Figure 3: MapReduce algorithm for Relation Filtering and Aggregation.

Since COHeRE aims to audit ontologies utilizing external knowledge, \( (C_1, C_2) \) are discarded if they are multiply related in a single source vocabulary, that is, there exist \( i \) and \( j \) such that \( R_i \neq R_j \) but \( S_i = S_j \). For example, in the source vocabulary SNOMEDCT_US, concepts Acetaminophen (C0000970) and Acetaminophen 160mg/5mL elixir (C0973757) are related through two relationships: CHD (is a) and RO (has active ingredient). Such inconsistencies are discarded at the single source level. The relation aggregation process is implemented as a MapReduce Reducer in...
In sum, concept pair \((C_1, C_2)\) with relationship \(R\) and relationship attribute \(RA\) is filtered out if one of the following criteria is met:

- the relationship \(R\) is not \(CHD, RN, SIB, SY, AQ\) or \(RO\);
- the two concepts \(C_1\) and \(C_2\) share the same unique identifier \(CUI\);
- the relation attribute \(RA\) is mapped from or mapped to;
- if the two concepts \(C_1\) and \(C_2\) are multiply related in a single source vocabulary after relation aggregation.

### 2.2 Inconsistent Relation Detection

<table>
<thead>
<tr>
<th>No.</th>
<th>Concept pair ((C_1, C_2))</th>
<th>Relationships and source vocabularies ({(R, RA, S)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arginine supplement (C3853287), Arginine and glutamine supplement (C3853286)</td>
<td>(CHD, is a, SNOMEDCT_US)</td>
</tr>
<tr>
<td>2</td>
<td>Hexosamines (C0019478), Fructosamine (C0060765)</td>
<td>(CHD, null, MSH \mid CHD, null, NDFRT)</td>
</tr>
<tr>
<td>3</td>
<td>17-Oxosteroid (C0000167), Androstenedione (C0002860)</td>
<td>(SIB, null, CPM \mid CHD, null, MSH \mid CHD, null, NDFRT)</td>
</tr>
</tbody>
</table>

Table 3: Examples of concept pairs after the Relation Filtering and Aggregation process.

For each pair of concepts obtained in the Relation Filtering and Aggregation step, the concepts are related either in a single source vocabulary or in multiple source vocabularies. For instance, the concepts Arginine supplement (C3853287) and Arginine and glutamine supplement (C3853286) are related through \(CHD\) only in a single source vocabulary SNOMEDCT_US (No. 1 in Table 3); the concepts Hexosamines (C0019478) and Fructosamine (C0060765) are related through \(CHD\) in both MSH and NDFRT (No. 2 in Table 3); the concepts 17-Oxosteroid (C0000167) and Androstenedione (C0002860) are related through \(SIB\) in CPM and \(CHD\) in both MSH and NDFRT (No. 3 in Table 3).

![Figure 4](image)

Figure 4: Decision tree for inconsistent relation detection using the example concept pairs provided in Table 3 as the input. The numbers in the curly brackets indicate the row numbers in Table 3.

Concepts pairs related in a single vocabulary are considered consistent (see the decision tree in Figure 4). For each concept pair with multiple source vocabularies, the concepts are either related through homogeneous relationships (e.g., No. 2 in Table 3) or non-homogeneous relationships (e.g., No. 3 in Table 3). Concept pairs related through homogeneous relationships are considered consistent, while those related through non-homogeneous relationships are considered inconsistent and collected to facilitate further analysis.

### 2.3 Relation Voting

The inconsistent concept pairs detected above have non-homogeneous relationships across different source vocabularies. A voting module is developed to rank the non-homogeneous relationships by the number of claimed source
vocabularies. If the relationship with the most votes has at least 2 votes, then it will serve as a suggestion for correcting potential errors in the inconsistent source vocabularies; otherwise, no suggestion will be made by COHeRE. Take row 3 in Table 3 as an example, the CHD relationship receives 2 votes from the source vocabularies MSH and NDFRT, and the SIB relationship receives 1 vote from the source vocabulary CPM. Hence CHD relationship could be suggested to correct potential error or inconsistency in the source vocabulary CPM. The output of COHeRE are the inconsistent concept pairs and suggested relationships with the most votes.

3 Results

After the Relation Filtering and Aggregation step, 11,921,502 distinct concept pairs were obtained. Among these, 11,372,069 pairs were related in a single source vocabulary, and 549,433 pairs were related in multiple source vocabularies. After the Inconsistent Relation Detection step, 138,987 concept pairs were found with inconsistent relationships across multiple source vocabularies. Table 4 shows the inconsistent relationships containing at least one hierarchical relationship (e.g., CHD or SIB) ranked by the number of concept pairs involved. For example, there were 2,755 concept pairs with inconsistent relationships CHD and SIB, and 1,407 concept pairs with inconsistent relationships CHD and SY. After the Relation Voting step, the concept pairs are classified according to the number of highest votes. Table 5 shows the numbers of concepts pairs with the highest votes between 1 and 7 with specific examples provided. For instance, there were 962 concept pairs with highest votes at least 3; 51 concepts pairs with highest votes at least 5; and there were no concept pairs with highest votes greater than 6.

<table>
<thead>
<tr>
<th>Inconsistent Relationships</th>
<th>No. of Concept Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD, RO</td>
<td>22,644</td>
</tr>
<tr>
<td>SIB, RO</td>
<td>24,475</td>
</tr>
<tr>
<td>CHD, RN</td>
<td>6,199</td>
</tr>
<tr>
<td>CHD, SIB</td>
<td>2,755</td>
</tr>
<tr>
<td>CHD, SY</td>
<td>1,407</td>
</tr>
<tr>
<td>CHD, RN, RO</td>
<td>823</td>
</tr>
<tr>
<td>CHD, RN, SY</td>
<td>812</td>
</tr>
<tr>
<td>SIB (sib in part of), RO</td>
<td>797</td>
</tr>
<tr>
<td>SIB (sib in part of), SIB</td>
<td>433</td>
</tr>
<tr>
<td>CHD, SIB, RO</td>
<td>183</td>
</tr>
<tr>
<td>CHD, CHD (member of)</td>
<td>115</td>
</tr>
<tr>
<td>SIB (sib in branch of), SIB</td>
<td>72</td>
</tr>
<tr>
<td>CHD, SY, RO</td>
<td>52</td>
</tr>
<tr>
<td>SIB, SY, RO</td>
<td>46</td>
</tr>
<tr>
<td>CHD, SIB (sib in part of)</td>
<td>35</td>
</tr>
<tr>
<td>SIB (sib in branch of), RO</td>
<td>27</td>
</tr>
<tr>
<td>SIB, SY</td>
<td>25</td>
</tr>
<tr>
<td>SIB (sib in tributary of), SIB</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 4: Inconsistent relationships with at least one hierarchical relationship (CHD or SIB), ranked by the number of concept pairs involved. Here CHD represents CHD with relationship attribute is a, CHD without relationship attribute specified, or RN with relationship attribute is a; and CHD with other kind of relationship attribute specified is explicitly provided in this table such as CHD (member of). SIB represents SIB with relationship attribute sib in is a or SIB without relationship attribute specified; and SIB with other kind of relationship attribute specified is explicitly provided in this table such as SIB (sib in part of). RO represents RO with any relationship attribute or without relationship attribute specified. RN represents RN with any relationship attribute or without relationship attribute specified. SY represents SY with any relationship attribute or without relationship attribute specified.

3.1 Evaluation

Since hierarchical relationships are dominant in biomedical ontologies, this study focused on evaluating two types of inconsistent relationships [CHD, RO] and [CHD, SIB] (see Table 4). A random sample of 40 UMLS concept pairs with inconsistent relationships (20 for each type) was selected and manually reviewed by a human expert. Table 6 summarizes the 40 concept pairs into 7 subtypes with examples given. For the [CHD, SIB] type, source vocabularies include MDR, NCI, ICD9CM, ICD10CM, SNOMEDCT_US, GO, FMA, and NDFRT, and the examined 20 pairs of concepts are regarding to disease, syndrome, injury, neoplastic process, cell component, pathologic function, or biologically active substance. For the [CHD, RO (has ingredient)] type, source vocabularies include NDFRT, RXNORM,
<table>
<thead>
<tr>
<th>No. of Highest Votes</th>
<th>No. of Concept Pairs</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>7+</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>6+</td>
<td>2</td>
<td>Dengue Fever (C0011311), Severe Dengue (C0019100)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 votes for CHD: MTH, MSH, NCI, NDFRT, SNOMEDCT_US, MEDCIN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 vote for SIB: ICD10CM</td>
</tr>
<tr>
<td>5+</td>
<td>51</td>
<td>Methacholine (C0600370), Methacholine Chloride (C0079829)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 votes for CHD: MTH, MSH, SNOMEDCT_US, MEDCIN, NDFRT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 vote for RO (has free acid or base form): NCI</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 vote for RN (form of): RXNORM</td>
</tr>
<tr>
<td>4+</td>
<td>298</td>
<td>Thrombocytopenic purpura (C0857305), Thrombotic thrombocytopenic purpura (C0034155)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 votes for CHD: MSH, NCI, NDFRT, SNOMEDCT_US</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 vote for SIB: MDR</td>
</tr>
<tr>
<td>3+</td>
<td>962</td>
<td>Acetaldehyde (C0000966), Paraldehyde (C0030438)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 votes for CHD: MTH, MSH, NDFRT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 vote for SIB: CPM</td>
</tr>
<tr>
<td>2+</td>
<td>5,937</td>
<td>Orthostatic hypotension (C0020651), Postural Orthostatic Tachycardia Syndrome (C1299624)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 votes for SIB: MSH, MDR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 vote for CHD: SNOMEDCT_US</td>
</tr>
<tr>
<td>1+</td>
<td>146,571</td>
<td>Amylases (C0002712), Isoamylase (C0022143)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 vote for SIB: MSH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 vote for CHD: SNOMEDCT_US</td>
</tr>
</tbody>
</table>

Table 5: Numbers and examples of concept pairs with top votes at least \(n\) \((1 \leq n \leq 7)\). \(n^+\) means \(\geq n\).

\(\text{SNOMEDCT\_US}\), and \(\text{VANDF}\), and all the evaluated 12 concept pairs are regarding to clinical drug. For the \([\text{CHD, RO (regional part of)}]\) type, source vocabularies include \(\text{SNOMEDCT\_US}\) and \(\text{FMA}\), and all the 4 examined concept pairs are regarding to body part or organ component.

<table>
<thead>
<tr>
<th>Subtypes of Inconsistent Relationships</th>
<th>No.</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD, SIB</td>
<td>20</td>
<td>Ventricular Dysfunction (C0242973), Right Ventricular Dysfunction (C0242707)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHD: MSH, NCI, NDFRT, SIB: MDR</td>
</tr>
<tr>
<td>CHD, RO (has ingredient)</td>
<td>12</td>
<td>Buprenorphine / Naloxone (C1169989), Buprenorphine 8 MG / Naloxone 2 MG Sublingual Tablet (C1168831)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHD: SNOMEDCT_US, NDFRT; RO (has ingredient): RXNORM, VANDF</td>
</tr>
<tr>
<td>CHD, RO (regional part of)</td>
<td>4</td>
<td>Fifth lumbar vertebra (C0223552), Arch of fifth lumbar vertebra (C0223553)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHD: SNOMEDCT_US; RO (regional part of): FMA</td>
</tr>
<tr>
<td>CHD, RO (has finding site)</td>
<td>1</td>
<td>Ear structure (C0013443), Hyperacusis (C0034880)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHD: OMIM; RO (has finding site): SNOMEDCT_US</td>
</tr>
<tr>
<td>CHD, RO (negatively regulates)</td>
<td>1</td>
<td>Cell-Matrix Adhesion (C0887869), Negative Regulation of Cell-Matrix Adhesion (C1516324)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHD: NCI; RO (negatively regulates): GO</td>
</tr>
<tr>
<td>CHD, RO (has manifestation)</td>
<td>1</td>
<td>Dextrocardia (C0011813), Kartagener Syndrome (C0022521)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHD: MSH, SNOMEDCT_US; RO (has manifestation): OMIM</td>
</tr>
<tr>
<td>CHD, RO (has manifestation), RO (disease may have associated disease)</td>
<td>1</td>
<td>Aniridia (C0003076), WAGR Syndrome (C0206115)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHD: MSH, NDFRT, SNOMEDCT_US; RO (has manifestation): OMIM, RO (disease may have associated disease): NCI</td>
</tr>
</tbody>
</table>

Table 6: Subtypes and examples of the 40 concept pairs with inconsistent relationships examined.

For each given pair of UMLS concepts, the expert was required to perform the following procedures:

(a) Check if there are terms mapping to them in each source vocabulary involved, and keep records of identifier codes for each term identified in a source vocabulary.

(b) For each relationship between the two concepts, check if the relationship is indeed claimed in the corresponding source vocabulary.

(c) For the relationship with the highest votes as at least 2, choose “agree,” “disagree,” or “not sure” to rate if the
expert agrees that it can be used as the correct relationship for the concept pair to resolve inconsistency. For the case that there is 1 vote for each relationship involved, specify one relationship as the correct relationship or answer “not sure.”

Take as an example the UMLS concepts \textit{Osteosarcoma metastatic} (C0278512) and \textit{Extraskeletal osteosarcoma metastatic} (C0855050) related through \textit{CHD (is a)} in NCI and SIB in MDR:

(a) Terms mapping to the concept \textit{Osteosarcoma metastatic} (C0278512) include C7781 in NCI and 10031246 in MDR. And terms mapping to \textit{Extraskeletal osteosarcoma metastatic} (C0855050) include C8808 in NCI and 10015849 in MDR.

(b) The relationships \textit{CHD (is a)} and \textit{SIB} between the two concepts are indeed claimed in NCI and MDR, respectively.

(c) The expert specifies that \textit{CHD (is a)} is the correct relationship.

The intention of steps (a) and (b) was to verify if the detected inconsistent relationships for the concept pair indeed exist in the source vocabularies or are caused by the integration process of source vocabularies into the UMLS. They are essential for evaluating COHeRE’s ability to filter out noisy information the integration of UMLS may cause and detect valid inconsistent relationships across different ontologies.

It was verified that all the 40 concept pairs have mapping terms in the involved source vocabularies. For a total of 95 individual relations (note that one concept pair have multiple individual relations in different source vocabularies), only 4 of them were found not claimed in the corresponding source vocabulary. For instance, the \textit{is a} relation between \textit{Metformin} (C0025598) and \textit{Metformin hydrochloride 1000 MG Oral Tablet} (C0978482) was no longer active in SNOMEDCT-US, but still included in the UMLS. In general, 95.8\% (91/95) of the individual relations detected by COHeRE is valid.

For cases where the relationship with highest votes is at least 2, there were 19 concept pairs, among which 14 (73.7\%) concept pairs were agreed by the expert, 2 disagreed, and 3 not sure. For the case that there is 1 vote for each relationship involved, there were 21 concept pairs, among which, 19 concept pairs were specified a correct relationship by the expert, and 2 not sure.

Manual examination of the verified result indicates the following facts:

- \textit{ICD9CM, MSH, and MDR} sometimes classify concepts related through \textit{CHD} as \textit{SIB}. \textit{Dermatitis Herpetiformis} (C0011608) and \textit{Juvenile dermatitis herpetiformis} (C0152092) are related through \textit{CHD (is a)} in SNOMEDCT-US, and \textit{SIB} in ICD9CM; \textit{Neurites} (C0085103) and \textit{Axon} (C0004461) are related through \textit{CHD} in GO and FMA, and \textit{SIB} in MSH; \textit{Ventricular Dysfunction} (C0242973) and \textit{Right Ventricular Dysfunctions} (C0242707) are related through \textit{CHD} in MSH, NCI, NDFRT, and SIB in MDR.

- \textit{SNOMED_CT} and \textit{NDFRT} show inconsistencies regarding to relations between pharmacologic substances and clinical drugs. For instance, the pharmacologic substance \textit{Buprenorphine / Naloxone} (C1169989) and the clinical drug \textit{Buprenorphine 8 MG / Naloxone 2 MG Sublingual Tablet} (C1168831) are related through \textit{CHD} in SNOMED_CT and NDFRT, and \textit{RO (has ingredient)} in RXNORM and VANDF. However, the pharmacologic substance \textit{Potassium Chloride} (C0032825) and the clinical drug \textit{Potassium Chloride 1 MEQ/ML Oral Solution} (C0979640) are indeed related through \textit{RO (has active ingredient)} in SNOMED_CT, and through \textit{RO (has ingredient)} in NDFRT. This indicates that SNOMED_US and NDFRT are inconsistent for relating clinical drugs with pharmacologic substances.

- \textit{SNOMED_CT} sometimes relates body parts using \textit{CHD (is a)} instead of \textit{part of}. For instance, \textit{Fifth lumbar vertebra} (C0223552) and \textit{Arch of fifth lumbar vertebra} (C0223553) are related through \textit{CHD} in SNOMEDCT-US, and \textit{RO (regional part of)} in FMA.

4 Discussion

\textbf{Distinction from related work}. COHeRE leverages the knowledge across different ontologies to detect possible errors, which is distinct from most existing OQA works utilizing the knowledge provided within a targeted ontology [9, 10, 11, 12, 13, 14, 16, 17, 21]. COHeRE also differs from [22], where the structural disparity between SNOMED CT’s Body Structure sub-hierarchy and FMA was investigated using a cross-ontology method, but the scope was limited to ontological terms relating to human body structure. In [23], Mougin and Grabar exhaustively studied multiply-related concepts within the UMLS, and explored why the multiply-related concepts occur and whether they are inherited from source vocabularies or introduced by the UMLS integration. It was reported that a quarter of multiply-related
concepts in UMLS are inherited from source vocabularies [23]. COHeRE differs from [23] in two ways. One is that COHeRE adopts only a subset of source vocabularies that are most relevant for cross-ontology examination. The other is that COHeRE aims to achieve effective filtering mechanism to remove the multiply-related concepts caused by the UMLS integration, and utilize the actual inconsistent relationships across multiple source vocabularies to detect inconsistencies and facilitate ontology quality assurance.

Conceptualization difference analysis. Manual analysis showed conceptual difference between ontologies. Take the concept pair Nipple neoplasm (C1112166) and Benign nipple neoplasm (C1332519) as an example, COHeRE detected two inconsistent relationships: CHD (is a) in the source vocabulary NCI, and SIB in the source vocabulary MDR. In NCI, Nipple neoplasm is classified into Benign nipple neoplasm and Malignant nipple neoplasm. In MDR, “neoplasm” is classified according to “benign,” “malignant,” and “unspecified.” Nipple neoplasm is classified under Breast neoplasms unspecified malignancy, and Benign nipple neoplasm is classified under Breast neoplasms benign. Hence the detected inconsistency is due to the conceptual difference. Although concept terms containing “unspecified” were filtered out by COHeRE, other terms classified under such terms were still included for the inconsistency detection (e.g., Nipple neoplasm under Breast neoplasms unspecified malignancy). To avoid this, the hierarchical information provided by UMLS (the distribution file MRHIER) can be used to remove the descendant terms of those terms containing “unspecified.”

Limitations. First, this study relies on the UMLS knowledge source, and it is not generalizable to domains lacking an integrated ontological system. Second, this study is limited in the number of concept pairs with inconsistent relationships evaluated. Third, this study only focused on evaluating two types of inconsistent relationships ([CHD, SIB] and [CHD, RO]). Evaluating more types of concept pairs may reveal more interesting inconsistencies. Fourth, this study did not investigate concept pairs that are multiply related in a single source vocabulary. It would be interesting to study inconsistent relationships occurring in a single source vocabulary. Lastly, if the ontologies integrated in UMLS are given in the Web Ontology Language (OWL) and the relations are defined as mutually exclusive, then simple OWL inferencing should be able to detect these inconsistencies.

5 Conclusion

This paper presented a novel cross-ontology method, COHeRE, to detect possible errors and inconsistencies in hierarchical relationships across multiple biomedical ontologies. COHeRE leverages the UMLS knowledge source and the MapReduce cloud computing technique for systematic, large-scale ontology quality assurance work. COHeRE is effective in detecting inconsistent hierarchical relations among UMLS source ontologies for quality assurance. The effectiveness of COHeRE indicates that UMLS provides a promising environment to enhance qualities of biomedical ontologies.

References


[17] Cui L, Tao S, Zhang GQ. Biomedical Ontology Quality Assurance Using a Big Data Approach. ACM Transactions on Knowledge Discovery from Data. [In press]


Using Workflow Modeling to Identify Areas to Improve Genetic Test Processes in the University of Maryland Translational Pharmacogenomics Project

Elizabeth M. Cutting, BA1,*, Casey L. Overby, PhD1,*, Meghan Banchero, MS, MBA1, Toni Pollin, PhD1, Mark Kelemen, MD, MSc, MBA1, Alan R. Shuldiner, MD1, Amber L. Beitelshees, PharmD, MPH1

1Program in Personalized and Genomic Medicine and Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

Abstract

Delivering genetic test results to clinicians is a complex process. It involves many actors and multiple steps, requiring all of these to work together in order to create an optimal course of treatment for the patient. We used information gained from focus groups in order to illustrate the current process of delivering genetic test results to clinicians. We propose a business process model and notation (BPMN) representation of this process for a Translational Pharmacogenomics Project being implemented at the University of Maryland Medical Center, so that personalized medicine program implementers can identify areas to improve genetic testing processes. We found that the current process could be improved to reduce input errors, better inform and notify clinicians about the implications of certain genetic tests, and make results more easily understood. We demonstrate our use of BPMN to improve this important clinical process for CYP2C19 genetic testing in patients undergoing invasive treatment of coronary heart disease.

Introduction

At The University of Maryland Medical Center (UMMC), the Translational Pharmacogenomics Project (TPP) aims to implement routine pharmacogenetic-based dosing and drug selection1. The first project has been to use CYP2C19 genotype test results to tailor antiplatelet therapy to cardiac stent patients. In this project, patients can be prescribed clopidogrel or alternative anti-platelet agents (e.g., prasugrel or ticagrelor) after cardiac stent implantation, based on their CYP2C19 genotype. Patients with loss of function variants in CYP2C19 are less responsive to clopidogrel and may be more effectively treated with alternative anti-platelet agents. To incorporate the use of genetic test results into routine clinical care, a multi-step and multi-actor process is necessary. There are many critical and time-sensitive steps required to deliver these genetic test results from the testing laboratory to the clinician. These results inform clinicians’ decisions to prescribe and dose anti-platelet therapy to a patient based on their genotype. Therefore, this information must reach the clinician in a timely manner. Genetic test results are not only time-sensitive but can be difficult to understand by non-genetics-expert physicians2, 3. This complex process may benefit from modeling methods to understand and improve internal workflow.

Workflow modeling is a useful tool in the ever-changing and fast-paced world of healthcare. All tasks in a healthcare system require standard processes, from patient diagnosis to treatment delivery4. Workflow modeling in healthcare serves to clarify the current processes that are being used as well as to illuminate potential improvements to those processes. For example, our team has previously used workflow models to document the overall implementation process for the UMMC TPP program5 incorporating Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for CYP2C19 genotype and clopidogrel therapy5,6 into patient care.

There are many types of modeling languages and methods to illustrate clinical processes. Rad et al., identified requirements for healthcare process modeling languages: complexity, security and privacy, user understandability, optimized models for different purposes, evolution of processes, and nested processes and integration7. They also determined that, although no one modeling language will completely encompass and satisfy all the needs of a healthcare process, each language is useful for different modeling purposes. Our choice to use the Business Process Model and Notation (BPMN) stems from the need to

*These authors contributed equally to this manuscript.
understand interactions between multiple actors in our process. The process of delivering genetic test results to the clinician requires coordination by and interaction between at least five different groups of actors or entities: the testing laboratory, clinicians, the call center, the TPP research team, and the Electronic Health Record (EHR). Some of these actors or entities contain sub-actors, for example, departments of an organization. A BPMN, when applied to healthcare processes, helps us understand internal procedures in a graphical notation and communicate these procedures in a standard manner. Furthermore, we chose BPMN for its depiction of end-to-end flow of an entire process, which allows us to understand all interactions in the process.

Other studies have chosen workflow modeling as a method for assessment and improvement of healthcare processes. Shiki et al., demonstrated the ability to use Universal Modeling Language (UML) modeling to “clarify functions and extract typical processes” within a hospital-based cancer registry. After conducting interviews and observational surveys, the authors created process models using activity, use-case, and class diagrams. They concluded that modeling helps to analyze hospital-based cancer registration in order to identify specific departmental tasks and decrease task workload. Similarly, Askari et al., used UML activity diagrams to help improve a health care program in an elderly patient fall management case study. This program coordinated nine separate facilities with the aim of creating universal care for fall-prone patients. Activity diagrams were created from interviews with stakeholders that elucidated “essential sub-processes”, “information/data sources involved”, and “organizational aspects” of processes such as risk profiling and fall prevention intervention. The authors concluded that their modeling method allows them to describe and visualize patient care development, as well as identify shortcomings and propose solutions for program management.

To learn about and model the process of delivering genetic test results to clinicians, we conducted focus groups with clinicians with experience ordering the CYP2C19 genotyping test and the research team for the TPP Project. The goal for our focus groups to learn about processes for delivering genetic test results to clinicians is distinct from the our previous work using workflow models to document the overall implementation process for the UMMC TPP program in Shuldiner et al. With the information gained from focus groups in this study, we used BPMN modeling to identify areas to improve organizational and technical processes supporting the delivery of genetic test results in the TPP project. Furthermore, our work provides recommendations for the improvement of these processes.

Methods

Three two-hour focus groups were conducted in order to gain information and insight into the current process of delivering genomic test results from the testing laboratory to the clinician in two personalized medicine projects: the TPP and Personalized Diabetes Medicine Program (PDMP). Our first focus group comprised clinical research coordinators involved in both projects. The second and third focus groups involved clinicians from cardiology (for TPP) and endocrinology (for PDMP). Here, we describe findings from our first and second focus group that provide insight into TPP genetic testing processes specifically. Questions asked during the focus groups were related to three topics of interest: documenting genetic test results, notifying clinicians of those results, and viewing the genetic test results. Scripts for focus groups were developed with input from the research team, including members of personalized medicine projects to ensure appropriateness of questions for participants. Focus group sessions were audio recorded, transcribed, and coded by two authors using Dedoose, a qualitative research analysis platform. Our codebook contains terminology relevant to the three topics of interest (documenting, notifying and viewing genetic test results). The research team also decided to include codes relevant to user experience design (usability, missing functionality, problems and proposed solutions, and desired features). With these data, a BPMN model was created to document the current genetic testing processes. These models were used to capture the steps by which a genetic lab report moves from the testing laboratory to the clinician. To verify accuracy, provide completeness, and improve validity of these models, member checking was performed.
Results

Focus Group Participants and Setting

Focus groups were conducted between December 2014 and February 2015 at UMMC. Our first focus group had a total of four participants: one laboratory professional and three clinical research coordinators. The second focus group had five participants: three cardiology fellows and two nurse practitioners. All participants had previous experience with CYP2C19 test results.

Findings from Focus Groups: Current process could be improved to reduce errors and decrease workload

We found that the most important issue brought up in our first focus group with TPP research team members was that all parts of the current process are labor-intensive and potentially error-prone, with particular impact on research coordinators and testing laboratory personnel. Participants provided comments, identified problems and proposed solutions regarding the CYP2C19 testing workflow process (documenting, notifying, viewing). Additionally, the second focus group, comprising physicians and nurse practitioners, identified improvements with respect to clinician education, laboratory report design, and notification methods.

Documenting genetic test results in a lab report was the first topic covered with focus group participants (Table 1). We found that entering information into the laboratory report can be error-prone because the information must be entered manually multiple times throughout the process. Genetic testing is often performed in laboratories outside of the typical hospital laboratory. And, laboratory systems may be separate from the hospital EHR. This can lead to input errors, potentially impacting patient care. Another issue was the necessary availability of a credentialed laboratory director to sign off on a report. Within the current process, this person must manually sign off on all reports before they can be sent to the clinician and made available in the EHR. This task puts stress on the laboratory director, who may have other constraints on his/her time.

To address these issues, participants suggested that establishing electronic data transfer between testing laboratory and EHR clinical systems could help eliminate errors. Such data transfer could allow for reports prepared by laboratory professionals and documented in the lab system to also be resulted in the EHR and made available to clinicians in fewer steps. Ideally, such an interface between clinical systems would also allow a laboratory director to remotely sign off on reports. From the clinician perspective, we also identified a need to document additional information in order to better orient clinicians who infrequently order the CYP2C19 test. Specific recommendations were to: (a) include more information about the genetic test on the lab report such as information about the specific test and (b) provide easier access to relevant external resources and references to corroborate their treatment decision. More specifically, participants recommended the inclusion of references to current clinical studies about CYP2C19 to help make their treatment decision.

<table>
<thead>
<tr>
<th>Topics Covered</th>
<th>Issues</th>
<th>Proposed Solutions</th>
</tr>
</thead>
</table>
| Documentation of information in genetic test results in lab reports | • Information is inputted multiple times throughout the process  
• Documentation process is labor-intensive  
• A credentialed laboratory professional must be physically present to approve the report  
• Information included in the lab report is insufficient | • Access from the research genetic laboratory to the hospital via Laboratory Information Management System (LIMS). This will allow laboratory professionals to input information into this system and generate a report. The LIMS will then automatically fax it to the credentialed lab professional and, potentially, directly transport the report into the Electronic Health Record (EHR).  
• Provide a graphical representation of results in the lab report (e.g., table or chart)  
• Provide easy access to relevant electronic resources, literature, and clinical guidelines. |
The next section covered was the method to notify clinicians of test results (Table 2). Currently, there is no notification of results within the EHR. Participants proposed that an alert message in the EHR is the preferred method of notification. Effective clinical decision support (CDS) can help notify an ordering clinician when genetic information relevant to the drug is available. And, conversely, when the lab result is entered into the system, CDS can help identify when there is a potential mismatch between the patient’s genotype and a previously ordered medication. Participants also noted that the degree of urgency should be distinguished when clinicians are notified of results. Patients with percutaneous coronary intervention (PCI) often stay in hospital for less than 24 hours following the procedure. And, once anti-platelet medications are prescribed in the outpatient setting, ambulatory cardiologists and primary care physicians are less likely to interrupt treatment to make a change. We also found that notification methods could be improved. Currently, the testing laboratory pages the ordering physician who then has to call back to receive the primary (verbal) result. In addition to this problem, participants explained that, many times, the patient is not currently under the care of the test-ordering physician, who receives this page. This adds extra steps to the process so that the current care team can be notified and made aware of the test results. Participants recommended that this page-and-call system be replaced by a “smart” pop-up that could alert a clinician about test results. One clinician participant described his/her preferred frequency for such an alert to be “once for a particular user and not every single time [that user] logs in”. Another proposed solution was to create a subscription system through which multiple care team members would be made aware of test results. Finally, there was agreement among clinician participants that secure healthcare email was the preferred notification method. They noted that one email could convey all of the complex information set in a convenient format that is easily accessible.

Table 2. Notification issues and proposed solutions identified during focus group discussions

<table>
<thead>
<tr>
<th>Topics Covered</th>
<th>Issues</th>
<th>Proposed Solutions</th>
</tr>
</thead>
</table>
| Methods of notification of available genetic test results | • Genetic testing often does not have discrete values that lend themselves to clinical decision support (CDS) alerts.  
• There is no active notification system in the EHR for this type of test result.  
• Critical test result notification systems are not tuned for genetic information. Some intermediate phenotypes resulted in excess contact to clinicians, even when the results did not change clinical practice.  
• Method of page-and-call is perceived as outdated and untimely  
• Over-alerting in the EHR remains a common complaint.  
• Identifying the correct physician for notification can be challenging in an area like interventional cardiology where physicians are often consultants to the primary care team, but primary care providers may not have experience with altering medications based on genetic testing. | • Implement an alert that is triggered when a physician begins the prescription ordering process for anti-platelet therapy that considers the patient’s genotype and the recommended drug and dosage.  
• Revise algorithm to discern “urgent” results from “non-urgent” results.  
• Email laboratory reports within a secured hospital system can relay more information and allow the clinician to read on their own schedule.  
• Implement an alert in the EHR that prompts the user once for a particular patient’s chart, when that chart is being viewed.  
• Provide a subscription system for notification of test results for the entire care team of a particular patient. |

The final topic covered, viewing genetic test results, is limited currently by the nature of the TPP Project (Table 3). Clinician participants expressed a desire to access the lab report and treatment recommendations in one document/screen. As a research project, however, treatment recommendations and interpretation of results are sent on behalf of the research team to the clinician. They are not available within the lab report.
We also found that the organization of the current laboratory report could be improved. Specifically, Clinician participants recommended a re-ordering of information in the report viewed in the EHR to include type of test ordered, results, and recommendations at the top of the page, with the supplementary information included underneath. Also, participants suggested that the visual appearance of the report to be rearranged to show a graphical representation of the test results to enable quick interpretation. When discussing the layout of the genetic test result, participants explained that the current report could be improved by placing the result of the test on the first page where it is viewed in the EHR. Currently, clinicians must apply an extra click to see the full report.

Table 3. Viewing issues and proposed solutions identified during focus group discussions

<table>
<thead>
<tr>
<th>Topics Covered</th>
<th>Issues</th>
<th>Proposed Solutions</th>
</tr>
</thead>
</table>
| How genetic test results are viewed in the EHR. | • As a research protocol, clinical recommendations are included in a document that is separate from the lab report and is not linked to report in EHR  
• Report is not visually appealing as viewed in the EHR  
• Navigation to the full report from the EHR requires multiple clicks | • When the TPP project transitions to a routinely-implemented clinical test, recommendations and clinical recommendations will be added directly into the report  
• Organization of information would ideally include the test ordered, the result, the genotype and the phenotype, and then the interpretation that is specific to the patient at the top of the page, before other information in the report  
• Show full report on first page of result section |

Workflow Model of Current Delivery Process of Genetic Test Results to Clinician

After analyzing qualitative data collected from TPP Focus Groups, we developed a BPMN model to document current genetic testing processes. Figure 1 illustrates these processes including the five actors and entities that work together to perform processes and complete the task. The five actors or entities include healthcare providers, the TPP research team, the patient medical records, UMMC laboratories, and the Translational Genomics Laboratory (TGL).

We found that certain redundancies appear in the process because there are multiple methods of notification. After the report is approved by the TGL laboratory director, there are four separate routes to deliver and notify clinicians of results (Figure 1A). For all CYP2C19 test results a paper copy of the test results are faxed to the UMMC clinical laboratory, manually entered into the EHR and can be accessed by a patients’ care team (Figure 1B), the TPP research team prepares a letter that includes both the test results and treatment recommendations and documents it in the patients’ paper chart, also accessible by the care team (Figure 1C). For actionable results, there are two additional routes of delivery and notification. Once test results are faxed to the UMMC laboratory call center center calls the ordering physician with actionable results (Figure 1D), the TPP research team also identifies the current care team and hand delivers the same letter that is documented in the patients’ paper chart (Figure 1E). A separate letter explaining test results is composed and mailed by the TPP research team to patients with actionable results (Not show in Figure 1).

Redundant delivery methods also lead to increased workloads for the TPP research team and unnecessary notifications for the clinicians. For example, added tasks are placed on the TPP research team in order to correctly identify the current care team of a patient. Figure 1F captures the issue of identifying the correct clinician to notify of CYP2C19 test results. In addition, only the ordering physician is notified by the UMMC laboratory call center of test results. Often, the ordering physician is not caring for the patient at the time results are available and is therefore incorrectly notified. The TPP research team has the added responsibility to identify and relay information to the correct care team members.
Discussion

Overall, we uncovered many areas for improving existing TPP genetic testing processes. Both focus groups raised similar issues. All participants agree that there are labor-intensive and potentially error-prone steps throughout the process, and insufficient support and resources to assist clinicians in treatment decisions. Furthermore, clinicians noted that the current report is not visually appealing or easily understood. We have identified clinicians’ needs from information gathered from focus groups. These needs can be addressed by the development of an improved workflow. First, clinicians need laboratory reports that are concise, easily understood and visually appealing in order to make clinical decisions based on test results. Second, clinicians also need to be able to access relevant electronic resources, literature, and clinical guidelines that will further aid in their decision. Third, we also found that a better process for making the laboratory report available to clinicians is necessary to reduce errors and the workload on laboratory technicians and directors. Added functionality, such as the ability of multiple clinicians to subscribe to a test result, is also necessary to alert all members of the care team of test results.

Our recommendations for solving these problems and meeting these needs include implementing clinical decision support (CDS) and establishing a Laboratory Information Management System (LIMS) capable of communicating (interfacing) with the EHR. With the addition of CDS in the form of alert messages and recommendations in the laboratory reports, clinicians will be more informed to make treatment decisions. The ultimate goal of our CDS recommendations is to notify the correct clinician and members of the care
team. These proposed improvements to the process have been combined to create a more optimal workflow model of the genetic laboratory result delivery process (Figure 2).

**Figure 2.** Proposed Workflow Model. Circles denote start and end of events, diamonds denote decision nodes, and rectangles represent tasks performed by the actors in that respective lane. Actionable results are defined as a patient genotype of “poor” or “intermediate” metabolizer. EHR, electronic health record; UMMC, University of Maryland Medical Center; LIMS, Laboratory Information Management System.

Clinical decision support has been included in the beginning of this process to automatically identify patients who are candidates for this type of genetic test (Figure 2A). Additional work is needed to establish automated approaches to identify patients eligible for CYP2C19 genetic testing. Previous work characterizing genetic information in US Food and Drug Administration (FDA) drug labels, for example, found there was little guidance for when to order a genetic test compared to guidance on how to use genetic test results once they are available. One approach has been to implement pre-test alerts recommending testing prior to prescribing were issued 1106 times for thiopurines and various drugs affected by CYP2C6 such as codeine.

We also recommend establishing a LIMS for the Translational Genomics Laboratory that can communicate with our EHR, allowing laboratory reports to be electronically transmitted from the LIMS and stored in the EHR (Figure 2B). Standards that set the language and structure of information are needed to ensure that information can be passed between the EHR and LIMS in a precise and unambiguous fashion. Most clinical systems and EHRs can send and receive clinical content as Health Level 7 (HL7) messages using HL7 and related standards. Establishing an HL7 interface between a TGL LIMS and the EHR would eliminate current steps to fax genetic test results from the TGL to the UMMC clinical laboratory, that are then manually entered into the EHR (Figure 1).

Further, a new LIMS has the potential to support linking with a knowledge base to augment information typically included in a laboratory reports. There is existing work, for example, developing the GenelInsight Suite to support laboratory and provider use of DNA-based genetic testing through the GI laboratory reporting tool for drafting informative reports on variants identified in patients, and the GI Clinic application for clinicians to access their patient’s genetic reports. It may also be possible to embed clinical
recommendations into lab reports. However, approval by an oversight committee would be needed (Figure 2C). Such an oversight committee might include individuals certified in clinical genetics, clinical cytogenetics, clinical molecular genetics, clinical molecular genetics and/or molecular genetic pathology (i.e., American Board of Medical Genetics and American Board of Pathology certifications) that are qualified to provide professional interpretation of test results.

Finally, CDS could potentially be included as an alert message to preemptively notify the care team that CYP2C19 results are available in the EHR to view (Figure 2D). An alert message, triggered by actionable CYP2C19 test results, for example might include both patient-specific results and treatment recommendations. There is existing work to implement CDS when genetic test results are available for patients. Studies have highlighted the need for point-of-care CDS to integrate clinically actionable results with the EHR\(^\text{15}\). Also, other studies have developed interruptive CDS alerts for at least eight genetic test results. These alerts are shown to be flexible with the types of information provided within them to allow for continual updating, as new information about the test in question is discovered\(^\text{16}\). Although the full laboratory report is also available, a pop-up notification with the necessary information will be more efficient for busy clinicians\(^\text{17}\).

**Conclusion**

By using BPMN to represent this complex process, we are able to see where crucial improvements can be made and where CDS could be implemented. BPMN serves as a learning tool to examine similar processes in other systems. This methodology allows for an effective modeling of a complex workflow. Through the identification of important actors, pathways, and problems in this workflow, an iterative improvement process can produce a more streamlined and efficient system.

Creating workflow models in the context of standard of care as we have here helps identify areas to improve processes. As the TPP project transitions from a study protocol to a routinely implemented clinical test, Figure 2 will serve as a document for discussing CYP2C19 testing processes without reliance on the TPP research team. This study is also part of a larger project to design and develop a software application to assist clinicians in making healthcare decisions based on genetic test results. These focus groups are the first in two rounds of focus groups. This first round has helped us to identify points in the current workflow that might be improved by such software. Where issues have been highlighted, improvements have been proposed for the current process of delivering genetic laboratory test results from the testing laboratory to the clinician in the TPP project at the University of Maryland Medical Center. Suggested improvements have the potential to reduce the number of actors in this process, improve the richness of the information transferred through the inclusion of decision support, and increase the overall efficiency of this process. These improvements are expected to improve patient care through the incorporation of genetic test results into routine clinical care, which will enhance physicians’ clinical decision making abilities and reduce the time between lab test completion and treatment delivery.

**Acknowledgments**

This work was supported by NIH U01HL105198 and AHRQ R21H5023390 grants. We thank Dr. Diana Hernandez (Columbia University), Ms. Rhea Cosentino (University of Maryland School of Medicine), Ms. Kathleen Palmer (University of Maryland School of Medicine), and Mr. Jean Fredo Louis (University of Maryland, College Park) for their contributions.

**References**

2. Overby CL, Cutting E, Williams MS. Using genomic medicine resources under three prescribing scenarios: issues with information seeking. Society of Behavioral Medicine Annual Meeting and Scientific Sessions (San Antonio, TX) 2015 (accepted).


A hybrid manifold learning algorithm for the diagnosis and prognostication of Alzheimer’s disease

Peng Dai1,2, PhD, Femida Gwadry-Sridhar1,3, PhD, Michael Bauer1, PhD, Michael Borrie4, MD, for the Alzheimer's Disease Neuroimaging Initiative*

1Department of Computer Science, The University of Western Ontario, London, ON, Canada; 2Robarts Research, London, ON, Canada; 3Lawson Health Research Institute, St. Josephs Healthcare, London, ON, Canada; 4Division of Geriatric Medicine, University of Western Ontario, London, ON, Canada

Abstract

The diagnosis of Alzheimer’s disease (AD) requires a variety of medical tests, which leads to huge amounts of multivariate heterogeneous data. Such data are difficult to compare, visualize, and analyze due to the heterogeneous nature of medical tests. We present a hybrid manifold learning framework, which embeds the feature vectors in a subspace preserving the underlying pairwise similarity structure, i.e. similar/dissimilar pairs. Evaluation tests are carried out using the neuroimaging and biological data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) in a three-class (normal, mild cognitive impairment, and AD) classification task using support vector machine (SVM). Furthermore, we make extensive comparison with standard manifold learning algorithms, such as Principal Component Analysis (PCA), Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and isometric feature mapping (Isomap). Experimental results show that our proposed algorithm yields an overall accuracy of 85.33% in the three-class task.

Introduction

Alzheimer’s disease (AD) is a genetically complex, progressive and fatal neurodegenerative disease, which leads to memory impairment and other cognitive problems1. There currently is no effective treatment that delays the onset or slows the progression of AD. According to the 2010 World Alzheimer report, there are an estimated 35.6 million people worldwide living with dementia at a total cost of over US$600AD billion in 2010, and the incidence of AD throughout the world is expected to double in the next 20 years2. Therefore, there is an increasing demand for diagnosing and treating AD. This paper focuses on the application of machine learning algorithms for Alzheimer’s disease diagnosis.

The diagnosis of AD is complex involving a variety of medical tests. These may include physical exam, neurological exam, mental status tests, and brain imaging, which will be used as a whole to assess memory impairment, judge functional abilities, and identify behavior changes, while ruling out other conditions that may cause symptoms similar to Alzheimer's disease. Current diagnosis of Alzheimer's relies largely on assessments of cognition and behavior, e.g. memory loss and behavior disorder, which start to decline in the later disease stages1,3. It is very difficult to treat neurological damage in current medical technology. Therefore, early detection of AD is of vital importance. Early detection of AD allows early treatment and thus help to delay the onset of AD symptoms, which could greatly improve the daily livings of patients.

There are many challenges in the early detection of AD. For example, it is very difficult to distinguish it from other types of dementia that begin in middle age, such as Pick disease. Another important thing that has to be mentioned is feature selection. A large amount of factors (e.g. biomarkers, metal status tests, etc.) contributes to the diagnosis of AD. It is very difficult to obtain the optimal set of discriminant features, which can best identify Alzheimer’s disease patients from healthy people or further classify different stages of AD. In this paper, we propose a hybrid manifold learning framework to embed multivariate feature vectors into a manifold, which preserves the pairwise similarity relationship. Some preliminary results are given. Generally speaking, our proposed algorithm consists of two parts, metric learning and manifold learning. Firstly, we utilize metric learning algorithm, i.e. Probabilistic Global Distance Metric Learning (PGDM)15, to construct a similarity matrix, which is then passed to isometric feature mapping (Isomap)13 to construct a manifold, which shows better discriminant property for Alzheimer’s disease recognition (i.e. disease diagnosis).

*Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf
The remainder of this paper is structured as follows. In the Background section, we present a general review of the application of machine learning algorithms in Alzheimer’s disease diagnosis. Then, we narrow down to manifold learning algorithms. In the Methods section, we describe our proposed approach in detail, together with necessary supporting information. Detailed experiment setup is also given in this section. In the Results and Analysis section, extensive evaluation tests are presented to show the performance of our method. Furthermore, comparison is made with a wide range of modern manifold learning algorithms, i.e. as Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and isometric feature mapping (Isomap). Finally, we draw conclusion and discuss about future work in the last section.

Background

Conventional manifold learning refers to “nonlinear dimensionality reduction methods based on the assumption that [high-dimensional] input data are sampled from a smooth manifold” so that one can embed these data into the [low dimensional] manifold while preserving some structural (or geometric) properties that exist in the original input space. Manifold learning is usually based on certain objective function. For example, PCA is designed to find the subspace with each dimension pointing to the direction with largest variance. By designing an appropriate objective function, it is possible to find a suitable manifold for AD diagnosis.

A large amount of work has been done on the application of manifold learning algorithms in the diagnosis of Alzheimer’s disease. The general idea is to design a manifold, which possesses more discriminant power in terms of classification tasks. Based on the whether a training set is needed, those approaches can be divided into two broad categories, i.e. unsupervised approach and supervised approach. For unsupervised algorithms, principal component analysis (PCA) was applied to the ADNI dataset for the early detection of AD in Lopez et al.’s work. Yang and coworkers utilized independent component analysis (ICA) to the MRI data in 2011. In 2011, Park proposed to use Multidimensional Scaling (MDS) to discriminate shape information between AD from normal controls. For supervised approach, it has begun to draw more attention due to its promising performance. It is usually formulated as prior information about the patient clusters, e.g. similar/dissimilar training set, which is similar to Xing et al.’s algorithm. Examples of this category are Wolz et al.’s neighborhood embedding approach and Shen et al.’s sparse Bayesian learning approach.

The progressive structural damage caused by AD can be noninvasively assessed by using magnetic resonance imaging (MRI) to measure cerebral atrophy or ventricular expansion. Therefore, many researchers have been focusing on MRI images, leading to massive application of image processing algorithms. Keraudren et al. proposed to localize the fetal brain in MRI using Scale-Invariant Feature Transform (SIFT) features. Suk et al. proposed a deep learning-based feature representation with a stacked auto-encoder, which combines latent information with the original low-level features.

Recently, researchers begin to use hybrid manifold learning algorithm to improve system performance. For example, Gray et al. reported to use the random forest algorithm together with MDS for multi-modal classification of Alzheimer’s disease in 2013. The algorithm introduced in this paper belongs to this category. Manifold learning algorithms are usually based on certain similarity metric. For example, the classical MDS was based on the distance matrix derived from L2 norm. Therefore, designing a proper similarity metric is of vital importance in the algorithm (Isomap for our proposed algorithm, and MDS for Gray et al.’s algorithm). Gray et al. improved the algorithm by utilizing the random forest algorithm as the first step to generate a better similarity metric, which is then passed to MDS for manifold construction. Random forest is based on decision tree, which is a ‘general purpose’ classification algorithm. The objective function is to maximize the classification accuracy but not to improve the similarity performance. In our proposed algorithm, we utilize Xing et al.’s metric learning algorithm to obtain an optimal Mahalanobis distance based on pairwise similarity prior knowledge. In other words, the learned distance yields small value for similar pairs and large value for dissimilar pairs, which is the optimization criteria of the algorithm. On the other hand, Isomap is developed based on the similarity graph of the data, which means the performance is highly dependent on the similarity measure. Therefore, an extra optimization step for optimal similarity measure is a straightforward approach to improve system performance.
Methods

In order to enact the model, we follow the description of our proposed algorithm below in Figure 1.

Figure 1. Schematic overview of the proposed methodology.

Our proposed algorithm is developed based on Isomap. In contrast to other nonlinear dimensionality reduction algorithms, Isomap can efficiently computes a globally optimal solution. Moreover, Isomap is based on the similarity graph calculated using certain similarity metric, which can be perfectly combined with metric learning algorithm. It’s because the direct input of Isomap is the optimal metric, which is different from many other algorithms. For example, PCA looks for the direction for maximum variance span of the high dimensional data, and ICA finds independent directions of the input data.

In classical Isomap, the Euclidean or Mahalanobis distance is used to construct the similarity graph, which is then used to construct the manifold. Therefore, an appropriate similarity measure is of vital importance. The probabilistic global distance metric learning (PGDM) algorithm is utilized to learn a robust metric, which is then used to construct a neighborhood graph. Then, the Isomap algorithm is utilized to learn the underlying global geometry of the data set. Finally, a support vector machine (SVM) classifier is trained for a three-class classification task.

Similarity metric learning

Distance metric learning aims to learn a distance metric (parameters) for the input data space from a collection of similar/dissimilar points that preserves the distance relation among the training data. In our present implementation, the probabilistic global distance metric learning (PGDM) algorithm is adopted. It is designed to learn a Mahalanobis distance which satisfies,

\[ g(A) = \sum_{(x_i,x_j) \in S} \|x_i - x_j\|^2_A - \log \left( \sum_{(x_i,x_j) \in D} \|x_i - x_j\|^2_A \right). \]  

(2)

where \(\|\cdot\|_A\) stands for the Mahalanobis distance; \(A\) is the Mahalanobis distance matrix; \(S\) and \(D\) are the similar and dissimilar training sets, respectively.

In our present implementation, we only study a simplified scenario, where \(A\) is diagonal. Then, Equation (1) can be solved by using the equivalent form,

\[ g(A) = \sum_{(x_i,x_j) \in S} \|x_i - x_j\|^2_A - \log \sum_{(x_i,x_j) \in D} \|x_i - x_j\|^2_A. \]  

(2)

We can thus use Newton-Raphson method to efficiently solve Equation (2)15.

It has to be noted that the Mahalanobis distance gives different weight to different dimensions, which is equivalent to feature selection. Lower weight indicates lower impact on the final result, i.e. similarity measure.

Manifold learning

The isometric feature mapping (Isomap) is a nonlinear dimensionality reduction method, which is also a low-dimensional embedding method. Isomap is used for computing a quasi-isometric, low-dimensional embedding of a set of high-dimensional data points. The algorithm provides a simple method for estimating the intrinsic geometry of a data manifold based on a rough estimate of each data point’s neighbors on the manifold.
Figure 2. Embedding multivariate medical records into a manifold.

As shown in Figure 2, the raw feature vectors are firstly transformed into similarity distance matrix (equivalent to neighborhood graph). According to the Isomap algorithm, the traditional MDS is then applied to the matrix to construct a manifold.

SVM Classification

In our present implementation, a support vector machine (SVM) classifier is utilized. It has to be noted that our focus is the construction of a better manifold which is more suitable for diagnosis. SVM is only implemented for demonstration purposes, which can be replaced by other algorithms, such as Neural Networks, nearest neighbor, etc.

An SVM is a high-dimensional pattern classification algorithm, which constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. The data points are mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. SVM can be formulated as

\[ \arg \min_{(w,b)} \frac{1}{2} \|w\|^2 \]
\[ \text{s.t. } y_i (w_i - b) \geq 1 \]

where \( w \) is the normal vector to the hyperplane; The parameter \( \frac{b}{\|w\|} \) determines the offset of the hyperplane from the origin along \( w \).

Results and Analysis

In this section, we first discuss the dataset that was used in the analysis. Then, we present the detailed experimental results for our proposed algorithm.

Data acquisition and pre-processing

Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year publicprivate partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.
The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada22. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-222. For up-to-date information, see www.adni-info.org.

We obtained MRI image data and diagnosis information from the ADNI database. We applied the proposed methodology to neuroimaging and biological data from 822 ADNI participants (229 normal patient, 405 MCI patients, and 188 AD patients). Each patient has three different high dimensional vectors of medical test data, corresponding to three distinct visits to the hospital. For example, for each visit there is a $p$-dimensional vector, corresponding to all the tests (totally $p$) done at that time. All MRIs were sagittal T1-weighted scans. The scans were collected using a 1.5 T GE Signa scanner with an MR-RAGE acquisition sequence.

We excluded all invalid records (with missing feature entries), resulting in totally 2158 high dimensional data points. Direct recognition using the 2158 raw data points based on the MRI images is very challenging, since it is very difficult to identify importance features. Therefore, instead of using the raw images, the MRI images are preprocessed by the CIVET software to get the brain volume information, such as volume for 3rd ventricle, 4th ventricle, right/left brain fornix, right/left frontal, right/left globus pallidus, right/left occipital, etc. A complete list of features used in this paper is given at my personal website17.

Experiment setup

As described in previous section, there are totally 2158 data points, with 586 normal records, 1006 MCI records, and 416 AD records. We randomly select 50 normal, 50 CI, and 50 AD data points as the testing set, leaving the rest (i.e. 536 normal, 956 MCI, 366 AD) as the training set (2008 training points and 150 testing points).

The training/testing set preparation process is repeated 10 times, which leads to 10 training/testing set combinations. For each training/testing set, our proposed algorithm is applied to construct a manifold, which is then passed to a SVM classifier for a three-class recognition task (Normal, MCI, and AD). Therefore, there are totally 10 set of recognition results. We use the sample code at the corresponding authors’ website for Isomap and PGDM implementation23. Relative improvement is defined as

$$R_{im} = \frac{r_p - r_t}{r_t} \times 100\%$$

where $R_{im}$ is the relative improvement; $r_p$ is the recognition rate of our proposed algorithm; $r_t$ is the recognition rate of the comparison target.

Experimental results

Figure 3 gives the error bar plot of experimental results for different conditions, i.e. average of 3 classes, normal, MCI, and AD. In order to eliminate the bias caused by training/testing set partition, we take the mean and variance of the recognition results. The error bars are calculated from the mean and variance of the recognition results from ten training/testing sets.

It can be seen that as the dimension of the manifold increases the recognition rate increases significantly. For example, in the 3-class recognition task, shown in Figure 3(a), the recognition rate increases from 33% at 1 dimensional feature to around 83% at 27 dimensional (or more) feature. Furthermore, at lower dimension (e.g. $<10$), the change of dimension number gives more impact on the final result, while at high dimension (e.g. $>10$) the performance improvement resulting from feature dimension increase becomes very small. As feature dimension changes from 1 to 10, the recognition rate increases by around 40% (from 34% to 80%). However, as feature dimension (i.e. the number of features) further increases, the recognition rate increases only 4% with fluctuation.
Based on one’s objectives, the optimal dimension number may vary. For example, if the objective is to classify if the patient is MCI, the optimal feature dimension can be chosen to 10. However, if recognition accuracy is the main concern, the optimal feature dimension would be around 25. In our present work, the optimal dimension is chosen to be 15 for a balance between recognition rate and efficiency.

Another thing that has to be mentioned is that the recognition results for MCI are much better than Normal and AD. This is due to the fact that we possess more MCI training materials (956 feature vectors) compared to 536 normal feature vectors and 366 AD feature vectors. More training materials yield better recognition results, which are consistent with the results in Figure 3. The Normal training records are more than AD, corresponding to higher recognition rate for Normal. The amount of training materials can be arranged according to special application purposes.

![Figure 3: Recognition results for different conditions.](image-url)

Table 1 gives the confusion matrix based on the experimental results in Figure 3. It can be seen that for MCI patients our proposed algorithm yields a very promising recognition rate, 98.4%, which means high sensitivity. However, the
misclassification rate (false positive) is also very high, i.e. low accuracy. This means our propose algorithm tend to misclassify other category as MCI (1 – 67.86% = 32.14%). The physical meaning is that for a patient with MCI our proposed algorithm can probably recognized the patient as MCI (98.4%, high sensitivity). However, if our propose algorithm classify the patient as MCI, there is 32.22% chance that the diagnosis is incorrect. In this case, 10.76% chance the patient is actually Normal, and 21.38% chance the patient is AD. For the other two categories (Normal and AD), the false positive rates are low. However, the recognition rates drop to 84% and 68.8%, respectively. This means if the recognition result is Normal or AD, it is very likely (> 98%) that the diagnosis is correct. However, there is less chance the algorithm will go to this category i.e. low sensitivity. On average, the proposed algorithm shows better performance at classifying Normal (high sensitivity 84% and high accuracy 98.59%).

Table 1. Confusion Matrix, feature dimension is chosen to be 30.

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th>Rate (sensitivity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>MCI</td>
</tr>
<tr>
<td>Normal</td>
<td>420</td>
<td>78</td>
</tr>
<tr>
<td>MCI</td>
<td>5</td>
<td>492</td>
</tr>
<tr>
<td>AD</td>
<td>1</td>
<td>155</td>
</tr>
<tr>
<td>Rate (accuracy)</td>
<td>98.59%</td>
<td>67.86%</td>
</tr>
</tbody>
</table>

Comparison with other manifold learning algorithm

Extensive comparison is made state of the art manifold learning algorithms, i.e. PCA, MDS, and Isomap. Here, Isomap utilizes the L2 norm to construct the neighborhood graph. We also give the recognition result based on the original feature vectors, denoted as Original in Table 2. The raw data vectors are processed by mean and variance normalization before processed by each of the above mentioned algorithms. Table 2 gives the experimental results for all the comparison targets.

Table 2. Recognition results for comparison targets (%).

<table>
<thead>
<tr>
<th></th>
<th>Proposed</th>
<th>Isomap</th>
<th>PCA</th>
<th>MDS</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition accuracy</td>
<td>83.83</td>
<td>82.83</td>
<td>83.00</td>
<td>82.67</td>
<td>74.66</td>
</tr>
<tr>
<td>Relative Improvement</td>
<td>-</td>
<td>1.21</td>
<td>1.00</td>
<td>1.40</td>
<td>12.28</td>
</tr>
</tbody>
</table>

It can be seen that our proposed algorithm yields much better results than all the comparison targets. The relative improvements are 1.21% over Isomap, 1.00% over PCA, 1.40% over MDS, and 12.28% over Original. The detailed recognition results for different conditions are visualized in Figure 4. Our proposed algorithm shows consistently better results than all the comparison targets. Since we have more MCI training materials, the corresponding MCI results are much better than the others, followed by Normal and AD.
Conclusion and future work

In this paper we propose a general framework for embedding multivariate medical records into a low dimensional manifold which possesses better discriminative property. This proposed algorithm is designed by the combination of two algorithms. We first construct a neighborhood graph based on a robust similarity metric, which is then used for manifold learning. It has to be noted that the PGDM algorithm is used to learn a Mahalanobis distance, which assigns different weights to different dimensions. This is equivalent to performing feature selection on the whole feature set. For example, in our present, the 23th feature (i.e. right globus pallidus volume) is given very low weight, which indicates that right globus pallidus volume is not suitable for the proposed classification task.

Our present study focuses mainly on the diagnosis of Alzheimer’s disease. We utilize all the features as a whole for the recognition task. However, it would be better if we can identify which feature is more important in identifying AD. The application of feature selection algorithms will help to further improve the performance of the proposed algorithm. Furthermore, our present work cannot predict the future tendency of a patient. Risk analysis of a patient in terms of developing Alzheimer’s disease will be our next step.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education,
and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

17. https://sites.google.com/site/declanide/
Extracting Characteristics of the Study Subjects from Full-Text Articles

Dina Demner-Fushman, MD, PhD, James G Mork, MS
National Library of Medicine, National Institutes of Health, HHS Bethesda, MD, USA

Abstract

Characteristics of the subjects of biomedical research are important in determining if a publication describing the research is relevant to a search. To facilitate finding relevant publications, MEDLINE citations provide Medical Subject Headings that describe the subjects’ characteristics, such as their species, gender, and age. We seek to improve the recommendation of these headings by the Medical Text Indexer (MTI) that supports manual indexing of MEDLINE. To that end, we explore the potential of the full text of the publications. Using simple recall-oriented rule-based methods we determined that adding sentences extracted from the methods sections and captions to the abstracts prior to MTI processing significantly improved recall and F1 score with only a slight drop in precision. Improvements were also achieved in directly assigning several headings extracted from the full text. These results indicate the need for further development of automated methods capable of leveraging the full text for indexing.

Introduction

Retrieval of publications for clinical decision support or database curation often relies on information about study subject characteristics. In publications indexed for MEDLINE®, this information is provided in structured, normalized form in Medical Subject Headings® (MeSH®) and can be easily used to find, for example, articles pertaining to preschool children or inbred mice. When this information is not available in structured form or directly stated in MEDLINE abstracts, it needs to be derived or extracted from the full text of the publications. The U.S. National Library of Medicine (NLM) Medical Text Indexer (MTI) tool1 that assists manual annotation of MEDLINE citations provided by the NLM Index section is fairly accurate in extracting and deriving some of this information available in the abstracts. For example, MTI achieves 93.84% F-score on assigning the heading “Humans”, whereas some other headings pertaining to study subjects are still far from this level of accuracy2. Indexers at NLM have access to the full text of an article, while only the title and abstract are provided to MTI. Over the years, Indexers have noted that specific detailed information about the study subjects that MTI missed was included in the full text of an article, but, not in the title or abstract. In this work, we present simple methods for extracting specific Mice and Rat strains, species, age groups, and gender from the full text of publications and explore if information extracted from the full text improves MTI performance for these headings.

Extraction of the subjects’ characteristics has been addressed in the past; mostly separately for clinical trials and for identifying animal species studied in the paper. For clinical trials, researchers are extracting such characteristics as the names and time points for primary and secondary outcomes, eligibility criteria and sample size. In the most common approach to extraction of the trials characteristics, the sentences potentially containing this information are identified first using statistical methods, and then knowledge-based methods are used to extract information. Xu et al. found candidate sentences in MEDLINE abstracts using HMM classifiers and then used syntactic parse patterns and rules to extract subject descriptors, such as man, elderly, diabetic, brain-injured; the study size (number of participants) and the studied disease3. Note that Xu et al. did not further separate subject descriptors by age, gender, or disease class. Similarly, de Bruijn et al. used an SVM trained on 78 full text articles to identify the most promising sentences and then applied simple extraction patterns and rules (called weak extraction rules) to identify over 20 trial characteristics, such as the enrollment start date and the experimental and control interventions4. The weak rules implemented in the ExaCT system5, in which the extracted text for each of the 21 trial characteristics were assessed by curators, corrected and then stored in the database, were sufficient for identifying the majority of trial elements in sentences recommended by SVM classifiers. Zhao et al. used Maximum Entropy classifiers to first assign sentences in a collection of 19,893 medical abstracts and full text articles to one or more of five classes: Patient, Intervention, Result, Study Design, and Research Goal6. Then the words in the sentences assigned to these classes were further classified as: Sex, Condition, Race, and Age. Two of these characteristics overlap with our targets: gender and age. In a 5-fold cross validation on 52 words in the Sex class, and 175 words in the Age class, Zhao et al. achieved 90% F-score for gender and 81% F-score for age extraction. Similarly, Kelly and Yang report perfect recall and precision for regular expression-based extraction of gender and age from 17 MEDLINE sentences containing information about the subjects’ age and 171 sentences with gender information7.
Several systems for identification of species in biomedical text are publicly available. Gerner et al. have developed a dictionary-based system LINNAEUS that was evaluated using MeSH species headings in MEDLINE citations, among other reference standards8. When using MEDLINE abstracts as input to the system, LINNAEUS achieved 52% precision and 68% recall. On the full text of the open access subset of PubMed Central, the system achieved 95% recall with a significant drop in precision to 13%. Naderi et al. have developed a hybrid rule-based/machine learning system, OrganismTagger9. Pafilis et al. have developed an open source SPECIES tagger, comparable in performance and faster than LINNAEUS10. As part of a large-scale multi-level event extraction effort, Pyysalo et al. achieved over 86% F-score extracting organism mentions, among other entities, using a single model that jointly predicts all entity types.11

In the previous work directly concerned with improving MTI performance, Jimeno-Yepes et al. compared extraction of MeSH terms from the abstracts, full text articles, and automatic summaries of different lengths12. For the headings addressed in our current work, the authors found that the machine learning methods currently implemented in MTI and trained on the abstracts have higher precision and somewhat lower recall than when summaries or full text are used as input. In this work, we continue exploring if the benefit of high recall offered by the full text of an article could be leveraged without significant losses in precision. Since the rule-base methods have shown high recall and precision in the previous evaluations discussed above, we start our targeted exploration of the study subjects’ characteristics with rule-based methods.

Our goals are threefold: first, we still do not have conclusive evidence that full text will significantly improve MTI performance on the headings pertaining to subjects’ characteristics; second, we would like to know if focusing on specific sections of the articles, particularly the methods section or captions will be more beneficial; and third, we would like to know if we should augment the original citations with the sentences extracted from the full text and then process these augmented citations using the MTI algorithm, or if we should directly assign the headings to citations using manually prepared lists of mappings of the extracted characteristics to MeSH.

In this work, we explored extraction of the following 29 MeSH Check Tags: Adolescent; Adult; Aged; Aged, 80 and over; Animals; Bees; Cats; Cattle; Cercopithecus aethiops; Chick Embryo; Child; Child, Preschool; Cricetinae; Dogs; Female; Guinea Pigs; Horses; Humans; Infant; Infant, Newborn; Male; Mice; Middle Aged; Pregnancy; Rabbits; Rats; Sheep; Swine; and Young Adult. We also explored all 51 specific strains of Mice and Rats under the Murinae [B01.050.150.900.649.865.635.505] 2015 MeSH tree, collectively identified as subject terms in this study. The specific Mice and Rat strains were included in our study because they are also typically mentioned in the full text of a paper and not in the title or abstract. Check Tags are a special type of MeSH term that is required to be included for each article and covers species, sex, human age groups, historical periods, pregnancy, and various types of research support (e.g., Male)13.

Methods

We used the 2014 MTI Test Collection that contains 143,658 citations randomly selected from the pool of citations indexed in the last year14. Of these, 14,829 (10.32%) full-text articles are available in the Open Access subset of PubMed Central15. We downloaded the articles in XML format and used the XML structure to evaluate extraction of the subject terms from various sections of the full text.

To identify the Methods sections that are most likely to describe the study subjects, we first extracted all section headings from the XML files. We then manually reviewed the names of the sections and the section titles and created a lookup list of the section names and titles most likely pertaining to Methods.

We then implemented a simple one-pass algorithm that parses the XML files, identifies candidate sentences using trigger words and extracts subjects’ characteristics from the candidate sentences. When the XML structure is parsed, information about the current section is stored and assigned to the sentences extracted from the section.

In the first step, the algorithm identifies candidate sentences potentially containing the subjects’ characteristics. Aiming for high recall, we qualify a sentence to be a candidate if it contains any members of the lookup lists or matches subject-related regular expressions described below. Using the section label, we determine if the sentence is found in the title/abstract, methods, caption, or anywhere in the body of the paper, excluding the abstract.
In the second step, the algorithm applies the gender and age extraction rules and looks up a MeSH heading corresponding to the list entry or the regular expression found in the sentence. We established the mappings for each entry and expression manually as described below. Finally, using the section label attached to each sentence we generate the following files for our experiments: subjectLinesMethods.txt (sentences extracted from the Methods sections only), subjectLinesMethodsCaptions.txt (sentences extracted from the Methods sections and Figure and Table captions), and subjectLinesBody.txt (sentences extracted from any section in the paper).

The final output of the algorithm consists of the sentences and information extracted from the sentences as shown in Figure 1.

![Figure 1 MeSH headings extracted from full-text sentences. The pipe-separated output presents: PMID; section of the full text (B stands for Body); strains and species headings; gender headings; age headings; and the sentence.](image)

**Dictionaries and regular expressions**

For identifying study subjects, their gender and age, we adapt the dictionaries and algorithms developed previously to identify patients’ characteristics in MEDLINE abstracts and clinical text. Briefly, our lookup list for human study participants consists of the manually curated concepts in the UMLS semantic type Population Group. We expanded the study subject list with case-insensitive regular expressions corresponding to MeSH entry terms for the animals in our subject terms listed above. For example, for Swine, we added to the list the following terms:

```
\WPig[s]?\W, \WHog[s]?\W, Phacochoerus, Suidae, and Warthogs.
```

For gender, we use two case-insensitive regular expressions:

```
\W(male[s]?|man|men|boy[s]?)\W and \W(female[s]?|pregnan[tcy]?|women|girl[s]?)\W. The second expression also extracts the Pregnancy heading.
```

Finally, for the subjects’ ages, we used both a lookup list of the terms in the UMLS semantic type Age Group and MeSH Age Groups and a set of regular expressions for identifying exact subject ages and ranges, for example,

```
(?:mean|M)?\W*age[d|s]?\W*?(?range|from)?\W+\W*?(?to)?\W*\W*?\W*?(?year|day|week) or \W*\W*?\W*?\W*?\W*?
```

We normalized the exact ages to MeSH terms in our list of subject terms using MeSH Scope Notes, for example, ages >= 65 AND <= 79 map to Aged.

**Evaluation**

The current abstract-based MTI performance for Check Tags and the specific Mice and Rat strains serves as comparison in all our experiments and the actual human indexing for these citations serves as the reference standard.

To evaluate the contributions of the specific sections, we extracted sentences and headings from the methods sections alone, from the methods sections and captions, and finally, from anywhere in the body of the paper, excluding the abstract.
We evaluated the contributions of the full text under two conditions: 1) adding candidate sentences to the titles or abstracts of MEDLINE citations and processing these extended abstracts using the current MTI algorithm, and 2) normalizing the characteristics extracted from the sentences to MeSH terms and directly assigning these subject terms to citations.

We used recall, precision and $F_1$ score as evaluation metrics. We computed recall as the proportion of the gold standard subject terms that were correctly assigned by our tools and precision as the proportion of the subject terms assigned by the tools that were correct. The $F_1$ score is the harmonic mean of recall and precision.

Results

The corresponding XML tags consistently identified the abstract, captions and the body of the paper. We found significant variations in the section naming, both in terms of the XML structure and the titles themselves. The structure was either providing the section name as an attribute of the section tag: for example, `<sec id="Sec2" sec-type="materials|methods">` or providing the name as title after the section tag, for example, `<sec id="Sec13">Participants</title>`.

We collected 163 section type variations for the Methods sections, such as: "intro|methods", "materials", "materials-methods", "material|methods", "method", "methods", "methods|conclusions", "methods|results", "methods|subjects", "subjects", "subjects|methods" and 116 titles, such as "Methodology", "Methodology and Findings", "Methodology and Principal Findings", "Methodology/Findings", "Methodology/Principal Finding", "Methods", "Methods Findings", "Methods and Findings", "Methods and Results", "Methods and design", "Methods and materials", "Methods/Design".

Table 1 presents the results of the evaluation of full text either added to the titles and abstracts or directly contributing Check Tags from our list of subject terms. The type, which sentence file was used, Recall, Precision, $F_1$, and the number of Check Tags matched to the human indexing are provided for each experiment. Files used for the experiments include: subjectLinesMethods.txt (1), subjectLinesMethodsCaptions.txt (2), and subjectLinesBody.txt (3).

<table>
<thead>
<tr>
<th>Experimental set-up</th>
<th>File</th>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
<th>Matched CTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTI baseline (currently in use at NLM)</td>
<td>-</td>
<td>74.09%</td>
<td>81.35%</td>
<td>77.55%</td>
<td>31,588</td>
</tr>
<tr>
<td>Title expansion with sentences from the Methods section</td>
<td>1</td>
<td>78.19%</td>
<td>77.57%</td>
<td>77.88%</td>
<td>33,339</td>
</tr>
<tr>
<td>Abstract expansion with sentences from the Methods section</td>
<td>1</td>
<td>78.19%</td>
<td>78.26%</td>
<td><strong>78.20%</strong></td>
<td>33,315</td>
</tr>
<tr>
<td>Title expansion with sentences from the Methods section and captions</td>
<td>2</td>
<td>79.70%</td>
<td>76.05%</td>
<td>77.84%</td>
<td>33,983</td>
</tr>
<tr>
<td>Abstract expansion with sentences from the Methods section and captions</td>
<td>2</td>
<td>79.60%</td>
<td>76.79%</td>
<td>78.17%</td>
<td>33,937</td>
</tr>
<tr>
<td>Title expansion with sentences anywhere in the paper body</td>
<td>3</td>
<td>85.70%</td>
<td>58.07%</td>
<td>69.23%</td>
<td>36,541</td>
</tr>
<tr>
<td>Abstract expansion with sentences anywhere in the paper body</td>
<td>3</td>
<td>85.52%</td>
<td>59.86%</td>
<td>70.42%</td>
<td>36,463</td>
</tr>
<tr>
<td>Direct assignment of Check Tags with sentences from methods and captions</td>
<td>2</td>
<td>79.28%</td>
<td>74.17%</td>
<td>76.64%</td>
<td>33,802</td>
</tr>
<tr>
<td>Direct assignment of Check Tags with sentences anywhere in the paper body</td>
<td>3</td>
<td>86.42%</td>
<td>55.97%</td>
<td>67.94%</td>
<td>36,848</td>
</tr>
</tbody>
</table>

Sentences extracted from the full text consistently increased recall in assignment of Check Tags, independently of the way they were used. In all cases we also observed a drop in precision, however, for the sentences extracted from the Methods section and added to the abstracts (bolded $F_1$ in Table 1), the drop in precision was relatively small and the $F_1$-score has increased compared to the MTI baseline. As the scope of the included text increased, the numbers of citations for which candidate sentences were found also increased from 2,696 for 20,813 sentences from the methods sections only, to 9,834 for 79,610 sentences from the methods sections and captions, and to 326,993 sentences from 12,800 citations when the whole paper was considered. The fact that 2,029 articles, for which we had no candidate sentences, include Check Tags in the reference standard indicates that our lookup lists were incomplete.
The direct assignment of the Check Tags produced mixed results. Table 2 shows the results for the individual terms compared to the current MTI baseline. Only 13 of the 29 Check Tags showed moderate improvements offset by significant degradation in performance for the remaining Check Tags. Notably, for all of the Check Tags that rely on simply finding a term in the sentence, there was a significant drop in the F1 score. The gender tags are an exception to this observation with a slight improvement for both Female and Male tags. There were three cases (bolded results in Table 2) where Precision, Recall, and F1 all improved for a Check Tag (Aged; Aged, 80 and over; and Cricetinae). The age tags that rely on both the dictionary terms and patterns showed improvement for the terms that do not occur in the text often and have to rely more on extracting the ages and mapping the numeric values to headings. For example, given the sentence “Separate analyses were conducted for children (age 6-11y), adolescents (age 12-19y), and for younger (age 20-50y) and older adults (>= 51y)”, our algorithm extracts the following tags: Chick Embryo|B|6100|Humans| Yound Adult; Middle Aged; Adolescent; Adult; Child. While Child (children), Adolescent (adolescents), and Adult (older adults) are mentioned directly in the text, based on the MeSH age range rules, we also included Young Adult (19-24) and Middle Aged (45-64).

Table 2 includes Precision, Recall, and F1 for both the MTI Baseline and the Full Text Check Tag results and a column showing the differences between F1 scores. Improvements with Full Text are highlighted in tan, and major negative results are highlighted in yellow.

Table 2 Individual directly assigned tags compared to the current MTI baseline

<table>
<thead>
<tr>
<th></th>
<th>MTI Baseline</th>
<th></th>
<th>Full Text</th>
<th></th>
<th>F1 Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Adolescent</td>
<td>62.14%</td>
<td>28.89%</td>
<td>39.44%</td>
<td>56.24%</td>
<td>42.49%</td>
</tr>
<tr>
<td>Adult</td>
<td>66.34%</td>
<td>67.56%</td>
<td>66.95%</td>
<td>56.12%</td>
<td>73.02%</td>
</tr>
<tr>
<td>Aged</td>
<td>68.67%</td>
<td>59.76%</td>
<td>63.90%</td>
<td>68.70%</td>
<td>64.15%</td>
</tr>
<tr>
<td>Aged, 80 and over</td>
<td>49.82%</td>
<td>18.38%</td>
<td>26.85%</td>
<td>52.86%</td>
<td>26.22%</td>
</tr>
<tr>
<td>Animals</td>
<td>91.77%</td>
<td>82.54%</td>
<td>86.91%</td>
<td>86.19%</td>
<td>87.63%</td>
</tr>
<tr>
<td>Bees</td>
<td>75.00%</td>
<td>100.00%</td>
<td>85.71%</td>
<td>69.23%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Cats</td>
<td>62.79%</td>
<td>96.43%</td>
<td>76.06%</td>
<td>58.33%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Cattle</td>
<td>75.82%</td>
<td>79.31%</td>
<td>77.53%</td>
<td>69.65%</td>
<td>80.46%</td>
</tr>
<tr>
<td>Cercopithecus aethiops</td>
<td>60.00%</td>
<td>34.62%</td>
<td>43.90%</td>
<td>50.00%</td>
<td>40.38%</td>
</tr>
<tr>
<td>Chick Embryo</td>
<td>100.00%</td>
<td>10.53%</td>
<td>19.05%</td>
<td>83.33%</td>
<td>52.63%</td>
</tr>
<tr>
<td>Child</td>
<td>62.07%</td>
<td>55.99%</td>
<td>58.88%</td>
<td>42.42%</td>
<td>67.54%</td>
</tr>
<tr>
<td>Child, Preschool</td>
<td>70.30%</td>
<td>42.69%</td>
<td>53.13%</td>
<td>40.91%</td>
<td>51.37%</td>
</tr>
<tr>
<td>Cricetinae</td>
<td>56.41%</td>
<td>37.93%</td>
<td>45.36%</td>
<td>56.82%</td>
<td>43.10%</td>
</tr>
<tr>
<td>Dogs</td>
<td>84.68%</td>
<td>73.44%</td>
<td>78.66%</td>
<td>79.03%</td>
<td>76.56%</td>
</tr>
<tr>
<td>Female</td>
<td>82.37%</td>
<td>79.30%</td>
<td>80.81%</td>
<td>80.04%</td>
<td>85.35%</td>
</tr>
<tr>
<td>Guinea Pigs</td>
<td>94.44%</td>
<td>94.44%</td>
<td>94.44%</td>
<td>89.95%</td>
<td>94.44%</td>
</tr>
<tr>
<td>Horses</td>
<td>66.00%</td>
<td>97.06%</td>
<td>78.57%</td>
<td>79.77%</td>
<td>97.06%</td>
</tr>
<tr>
<td>Humans</td>
<td>89.86%</td>
<td>91.65%</td>
<td>90.74%</td>
<td>84.06%</td>
<td>93.01%</td>
</tr>
<tr>
<td>Infant</td>
<td>60.69%</td>
<td>45.69%</td>
<td>52.13%</td>
<td>41.98%</td>
<td>51.15%</td>
</tr>
<tr>
<td>Infant, Newborn</td>
<td>70.21%</td>
<td>41.60%</td>
<td>52.24%</td>
<td>30.00%</td>
<td>50.42%</td>
</tr>
<tr>
<td>Male</td>
<td>79.34%</td>
<td>78.41%</td>
<td>78.87%</td>
<td>77.34%</td>
<td>84.95%</td>
</tr>
<tr>
<td>Mice</td>
<td>92.96%</td>
<td>77.96%</td>
<td>84.80%</td>
<td>89.54%</td>
<td>87.56%</td>
</tr>
<tr>
<td>Middle Aged</td>
<td>74.75%</td>
<td>70.05%</td>
<td>72.32%</td>
<td>74.67%</td>
<td>73.91%</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>79.76%</td>
<td>88.50%</td>
<td>83.90%</td>
<td>66.73%</td>
<td>90.11%</td>
</tr>
<tr>
<td>Rabbits</td>
<td>89.23%</td>
<td>63.04%</td>
<td>73.89%</td>
<td>72.16%</td>
<td>76.09%</td>
</tr>
<tr>
<td>Rats</td>
<td>93.95%</td>
<td>74.69%</td>
<td>83.22%</td>
<td>88.61%</td>
<td>80.90%</td>
</tr>
<tr>
<td>Sheep</td>
<td>55.17%</td>
<td>88.89%</td>
<td>68.09%</td>
<td>20.37%</td>
<td>91.67%</td>
</tr>
<tr>
<td>Swine</td>
<td>74.03%</td>
<td>89.76%</td>
<td>81.14%</td>
<td>19.32%</td>
<td>93.70%</td>
</tr>
<tr>
<td>Young Adult</td>
<td>57.73%</td>
<td>20.00%</td>
<td>29.71%</td>
<td>56.03%</td>
<td>29.81%</td>
</tr>
</tbody>
</table>
Table 3 shows the results for 27 of the 51 specific strains of Mice and Rats where changes in performance were noted. The remaining 24 strains were not identified in the full text in this study and not included in Table 3. Similarly to the Check Tag performance in Table 2, we have mixed results for the specific strains of Mice and Rats as shown in Table 3. Only 12 of the 27 specific strains of Mice and Rats showed moderate improvements offset by significant degradation in performance for the remaining strains. Both “Rats, Inbred Lew” and “Rats, Zucker” (bolded results in Table 3) are cases where Precision, Recall, and F1 all improved from the use of full text. In the case of “Mice, Inbred SENCAR”, the results in Table 3 show no change in performance, but, in fact the full text provided 16 new cases of this term which were all incorrect.

Table 3 includes Precision, Recall, and F1 for both the MTI Baseline and the Full Text specific strains of Mice and Rats results and a column showing the differences between F1 scores. Improvements with Full Text are highlighted in tan, and major negative results are highlighted in yellow.

<table>
<thead>
<tr>
<th>MTI Baseline</th>
<th>Full Text</th>
<th>F1 Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>MH</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Mice, Inbred BALB C</td>
<td>61.36%</td>
<td>24.77%</td>
</tr>
<tr>
<td>Mice, Inbred C3H</td>
<td>50.00%</td>
<td>50.00%</td>
</tr>
<tr>
<td>Mice, Inbred C57BL</td>
<td>74.73%</td>
<td>28.96%</td>
</tr>
<tr>
<td>Mice, Inbred CBA</td>
<td>100.00%</td>
<td>30.00%</td>
</tr>
<tr>
<td>Mice, Inbred CTR</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Mice, Inbred DBA</td>
<td>100.00%</td>
<td>55.56%</td>
</tr>
<tr>
<td>Mice, Inbred ICR</td>
<td>50.00%</td>
<td>5.56%</td>
</tr>
<tr>
<td>Mice, Inbred NOD</td>
<td>77.78%</td>
<td>60.00%</td>
</tr>
<tr>
<td>Mice, Inbred SENCAR</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Mice, Knockout</td>
<td>69.11%</td>
<td>31.84%</td>
</tr>
<tr>
<td>Mice, SCID</td>
<td>76.00%</td>
<td>36.54%</td>
</tr>
<tr>
<td>Mice, Transgenic</td>
<td>76.60%</td>
<td>37.50%</td>
</tr>
<tr>
<td>Mice, Nude</td>
<td>71.15%</td>
<td>30.00%</td>
</tr>
<tr>
<td>Mice, WF</td>
<td>98.11%</td>
<td>98.11%</td>
</tr>
<tr>
<td>Mice, SCID</td>
<td>76.00%</td>
<td>36.54%</td>
</tr>
<tr>
<td>Mice, Nude</td>
<td>71.15%</td>
<td>30.00%</td>
</tr>
<tr>
<td>Mice, SENCAR</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Mice, Knockout</td>
<td>69.11%</td>
<td>31.84%</td>
</tr>
<tr>
<td>Mice, SCID</td>
<td>76.00%</td>
<td>36.54%</td>
</tr>
<tr>
<td>Mice, Transgenic</td>
<td>76.60%</td>
<td>37.50%</td>
</tr>
<tr>
<td>Rats, Inbred ACI</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rats, Inbred BB</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rats, Inbred BN</td>
<td>50.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Rats, Inbred BUF</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rats, Inbred F344</td>
<td>50.00%</td>
<td>25.00%</td>
</tr>
<tr>
<td>Rats, Inbred Lew</td>
<td>50.00%</td>
<td>20.00%</td>
</tr>
<tr>
<td>Rats, Inbred OLETF</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Rats, Inbred SHR</td>
<td>83.33%</td>
<td>83.33%</td>
</tr>
<tr>
<td>Rats, Inbred WF</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rats, Inbred WKY</td>
<td>60.00%</td>
<td>75.00%</td>
</tr>
<tr>
<td>Rats, Long-Evans</td>
<td>50.00%</td>
<td>28.57%</td>
</tr>
<tr>
<td>Rats, Sprague-Dawley</td>
<td>83.70%</td>
<td>37.93%</td>
</tr>
<tr>
<td>Rats, Wistar</td>
<td>83.05%</td>
<td>38.28%</td>
</tr>
<tr>
<td>Rats, Zucker</td>
<td>83.33%</td>
<td>83.33%</td>
</tr>
</tbody>
</table>

**Discussion**

Our first question – whether full text holds promise for improving MTI performance was answered positively, as we see significant improvements for thirteen Check Tags and twelve strains of Mice and Rats. The results also hint at the huge potential of full text with our highest recall of 86.42% compared to our MTI Baseline recall of 74.09%. Our results indicate however, that the simple methods that have been reported to show excellent results for extracting study subject characteristics lead to drop in precision for about half of the Check Tags and the specific strains of Mice and Rats. An example of such drop is the Check Tag *Horses* that relies on finding words *horse*, *horses* or *equus* in the text. It turns out, that the word *horse* occurs very frequently in the context of *horse serum* that...
is added as supplement to culture media. We will need to look at ways to improve our selection criteria and add filtering to reduce the false positives.

With respect to the second question, whether we should focus on specific sections of the articles, the differences in recall achieved when using the whole text and the Methods sections indicate that it might be more promising to focus on finding good candidate sentences anywhere in the text. Although the recall for candidate sentences was fairly high, our experiments indicated that our lookup lists are incomplete. Inspection of the full text of some of the articles from which no sentences were extracted showed that we systematically missed the age groups in the case reports for single patients because the terms man, farmer and woman were not included in our trigger terms. Generally, singular nouns were intentionally excluded from our trigger lists tailored for extracting characteristics of the subjects of clinical trials because for that task they often triggered false positives. Similarly, some age patterns are missing, for example, aged between \( d \) and \( d \) years. In some cases however, our algorithm will not be able to assign a Check Tag even if the trigger list is exhaustive. For example, the Check Tag Female was assigned to MEDLINE citation 23552690, however, the paper presents a study conducted on breast cancer cells and does not contain any gender-specific terms. It is quite possible that the specific cell lines discussed in the paper indicate that the tissues were female. If this assumption is correct, to assign the Check Tag to this article, we will either need to incorporate more domain knowledge or hope to find a sufficient number of examples for a machine learning algorithm. In the future, we will expand the lists using machine learning methods or manually inspecting citations that are indexed with Check Tags, but for which no candidate sentences were extracted using the current lookup lists. We also need to focus on identifying the higher quality candidate sentences to improve precision.

Our third question was whether we should add candidate sentences to the title or abstracts prior to MTI processing, or attempt to extract the subject terms directly using the simple mapping rules. The current results show that the answer for some of the age groups, strains of Mice and Rats, and gender might be to assign the subject terms directly, but for the majority of the tags augmenting the abstracts appears to be safer at the moment. Although the \( F_1 \) score is 1% higher than the current MTI baseline that supports the NLM Index section when the full text sentences are added to the abstracts, the corresponding 3% drop in precision indicates that we need to further explore how to use the full text.

Our work has the typical limitations of a feasibility study: we focused on testing the hypotheses rather than making sure that our trigger term lists are complete and all our extraction rules take into account the context surrounding the trigger terms. We chose to explore the hypotheses “breadth-first” exploring all parts of the full text, rather than following up “in-depth” with the very promising Methods sections results. These limitations also clearly define the future work that we plan to conduct shortly: expand the lists of trigger terms; maximize the benefits of using the Methods section; refine our extraction rules; and use the extracted sentences in weakly supervised machine learning experiments.

An additional consideration for pursuing methods based on the full text is its availability. MTI does not have access to the full text of an article at this time due to contractual reasons and only utilizes the title and abstract to produce its recommendations for the manual indexing performed at NLM. One possibility is to utilize the full text while it is available in memory for a relatively short period of time while it is being processed in the NLM Document Management System. To use the text in this short time, our algorithms need to be fast and at the same time offer significant benefits to justify the substantial efforts needed for including the full text in MTI processing. We hope that work like the research detailed in this paper will provide incentive for publishers to grant MTI access to full text in order to provide more complete recommendations for the MeSH indexing of their articles.

**Conclusion**

Our study shows that the full text of biomedical articles has potential to significantly improve automatic indexing of MEDLINE citations with MeSH headings pertaining to the study subjects’ characteristics. Furthermore, we show that simple rule-based methods significantly outperform the current automated indexing provided by NLM’s Medical Text Indexer for 25 of the 56 subject terms in our study, in some cases significantly better (Chick Embryo +45.47, Rats, Inbred Lew +42.86, and Mice, Inbred ICR +31.18). These encouraging results indicate we should continue exploring how to better use the full text for automated indexing of MEDLINE citations.
Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

PHAST: A Collaborative Machine Translation and Post-Editing Tool for Public Health

Kristin Dew, MS¹, Anne M. Turner, MD, MPH, MLIS², Loma Desai, MBA³, Nathalie Martin⁴, Adrian Laurenzi⁵, Katrin Kirchhoff, PhD⁶
¹-⁶University of Washington, Seattle, WA, USA

Abstract
This paper describes a novel collaborative machine translation (MT) plus post-editing system called PHAST (Public Health Automatic System for Translation, phastsystem.org), tailored for use in producing multilingual education materials for public health. Its collaborative features highlight a new approach in public health informatics: sharing limited bilingual translation resources via a groupware system. We report here on the design methods and requirements used to develop PHAST and on its evaluation with potential public health users. Our results indicate such a system could be a feasible means of increasing the production of multilingual public health materials by reducing the barriers of time and cost. PHAST’s design can serve as a model for other communities interested in assuring the accuracy of MT through shared language expertise.

Introduction and Background
There are more than 300 languages spoken in the US, with approximately 25.2 million people having limited English proficiency (LEP), defined as a limited ability to speak, read, write, or understand English.¹ For LEP individuals, accessing accurate and up-to-date health information can be difficult. LEP status contributes to poor health literacy and to a greater incidence of health disparities, such as less preventative health screening.²⁻⁵

As providers of health care, public health departments are required by law to provide language appropriate materials for individuals who do not speak English.⁶⁻⁸ This includes translating health education materials, such as handouts, web resources, and flyers, into various languages used in the departments’ respective communities. Translations are typically performed in-house by bilingual employees, or ordered from professional translation vendors. The high cost of translations, whether commissioned through professional vendors or using limited bilingual in-house staff, limits the number and type of language translations produced. These constraints are particularly true for smaller health departments with limited resources and staff.⁹ Unfortunately, the decentralized organization of public health in the United States means that local health departments often use their limited funds to translate content similar to that of other health departments, because they lack systems for sharing translation resources.

Over the last four years, as part of an NLM-funded research project, we have been investigating the potential role of machine translation (MT), i.e. the fully automatic translation of text or speech from one natural language into another by a computer program, in producing multilingual public health materials,¹⁰ and how MT could be integrated in current public health practice.¹¹ In this paper we describe the design and evaluation of PHAST, a collaborative translation tool developed to facilitate the integration of MT into the typical workflow of public health practice (phastsystem.org). An early prototype of PHAST was described previously.¹² This paper describes additional design work, the updated PHAST system, and an evaluation via user testing with public health staff.

Our target population comprises lay users – bilingual public health professionals who have a deep understanding of their area of public health or a particular community they serve, but no formal linguistic or technological expertise. In our initial studies,¹¹⁻¹³ we found that core characteristics required of any MT system for public health materials include: 1) a simple, intuitive user interface to facilitate the various steps involved without extensive training or maintenance; 2) support for MT quality control via post-editing, i.e. the manual review and correction of machine-translated text; 3) document sharing among different health departments; and 4) a means for users to track the progress of their translations.

The technical progress in MT over the past decade makes such a system possible. MT has improved dramatically in the past 10 years. With statistical MT – which is presently the most promising approach – models are trained automatically using large bodies of parallel text or text in the source language paired with its translation in the target language;¹⁴ a more detailed introduction to statistical MT technology can be found in Canciedda et al.¹⁵ Statistical MT has been shown to be more powerful than older approaches like rule-based MT and translation memories, which rely on sophisticated linguistic analyses or on large databases of stored examples that need to match the input text.
Therefore, the best statistical MT systems typically outperform rule-based or example-based MT systems, even on specialized technical text (see e.g., Bojar et al.). However, statistical MT alone is still too full of errors to be used reliably for end-to-end translation, especially in sensitive domains like health. In applications where high translation accuracy is required, statistical MT is combined with a human post-editing step, i.e. human readers correct the translation errors created by the MT engine. The MT plus post-editing approach has so far been under-utilized by local and regional health departments. Our previous work was the first to demonstrate the feasibility and advantages of using MT plus post-editing in the public health domain.

In a study of Spanish MT plus post-editing for public health materials, documents blindly rated as equivalent in quality to a professional, manual, human translation could be produced for 5% of the cost using MT plus post-editing by a bilingual public health professional, based on an average public health professional’s wage.

MT is now in common use amongst translation vendors and in various business applications, but has been slow to be adopted into clinical settings. MT followed by human post-editing, or correction to the machine-translated output, can hold substantial time and cost gains over manual human-only translation. Even with the required step of post-editing, our prior studies have shown that MT could help public health workers create multilingual documents in less time and at dramatically lower costs. In blind ratings by public health professionals, post-edited MT translations were found to be equivalent to manual human translations. In tandem with document sharing, MT plus post-editing could ultimately help health departments produce a greater number of low-cost multilingual health materials.

The barriers of time and cost are compounded by a lack of coordination across health care authorities. Our prior studies of Washington (WA) State health departments found that translated health materials generated at one health department were rarely shared among other health departments. While there is a wide range of bilingual language expertise across health departments (e.g. interpreters, translators, community health workers, etc.), no single health department has the language coverage to meet their translation needs. There have been studies of collaborative translation by monolinguals with machine translation systems, whereby two non-bilingual people who use different languages collaborate to perform translation using MT services. As a novel approach to translation and current public health practices, we designed PHAST to enable the remote collaboration of health department workers in editing translated documents produced through MT. To the best of our knowledge, the way in which language expertise and translated documents are shared in PHAST has not been tried until now.

We describe here the design and evaluation of the PHAST tool. While PHAST was developed as a public health informatics tool, we believe it contains useful ideas for designing novel technologies for lay users in other governmental domains, creating collaborative translation applications, and bootstrapping low-cost MT applications.

Methods
Design
User-centered design is essential for increasing the efficiency with which users interact with MT systems. We relied on several common user-centered design methods to develop PHAST. Its initial design requirements stemmed from information gathered with qualitative methods, including interviews and focus groups, within five health departments in Washington State (see Table 1). We used the Cognitive Work Analysis framework to model current translation workflow as part of a larger project investigating translation practices in Washington State public health agencies. This was the basis of the step-by-step translation workflow used in PHAST, as well as the scenarios used in its evaluation. The workflow study was supplemented with feedback from state health department staff, preliminary usability testing with lay users, and post-editing studies, forming an iterative design and evaluation process that culminated with the evaluation described below.

While some materials may be specific to the local health department, more general health education information – such as those on vaccines, basic food safety, and so on – are frequently produced independently by each health department and remain in the gray literature. We decided to extend the design requirements for PHAST to include collaboration on producing and storing translations, in the hopes of allowing health departments to pool their bilingual staffs’ language expertise and their translated documents (see System Overview below).

We used the findings from our studies of public health translation processes and those of the subsequent post-editing studies to develop a translation system that is easy to use, efficient, and flexible, and that pooled limited bilingual public health staff language resources. The system is designed for lay users rather than professional translators. As such, its design deviates from that of traditional translation management systems: based on the results of our workflow
studies, we removed irrelevant features and created a step-by-step translation workflow mirroring that of health departments’ current practices. Because the public health translation staff frequently cited time and cost constraints, we attempted to create a lightweight, simple and intuitive design, which would remove the need for extensive training and minimize adoption headaches for staff; deployment would simply consist of staff signing up and undergoing brief training, for which we have prepared a tutorial.

To minimize costs, we created PHAST as a web-based application built on freely available tools and requiring only minimal maintenance costs and technical resources. PHAST was built online using the Kohana PHP Framework and a MySQL database. The front-end interface was built using JQuery, Twitter Bootstrap, HTML and CSS.

PHAST supports the following four main tasks that our workflow studies found to be essential for users: (1) uploading a text document requiring translation from the source language (English), (2) statistical MT translation, (3) human post-editing of the MT output to ensure quality, and (4) saving the original and post-edited text documents in an archive available for users to download either the original or complete post-edited version (see Figure 1). An important constraint is that PHAST only handles raw text, so figures, bullet points, images, and other formatting must be removed before uploading a document to the system. Uploaded documents are visible in the main archive and can be sorted by date or language. Any user can see the document’s post-editing status, which is tracked with a progress bar that is updated as each line break is completed (see Post-editing Interface section). When all lines have been saved, the document is marked with a green check, indicating that it is available for download.

Figure 1: Overview of documents and their progress through post-editing

Our prior studies of WA State health departments found that, in general, health materials generated at one health department were rarely shared among other health departments. The PHAST system applies a collaborative approach to encourage sharing of bilingual resources across health departments. Public health workers using the system can upload and post-edit documents from within their own department and can volunteer to post-edit documents that have been uploaded by other departments. Therefore, most of the cost of translation stems from the cost of the post-editor’s time. Sharing bilingual staff expertise allows collaboration amongst different departments in translating and sharing health materials. PHAST pools both language expertise and translated documents with the features described below.

When a user creates an account, their agency affiliation, language expertise, and information about experience level and professional certifications are saved in a profile. Once a user has a profile, she can upload a document in PHAST, which uses the Microsoft Translator API. PHAST covers all 39 languages supported by the Microsoft Translator API. At present we are using a free license with a limit of 2,000,000 characters per month. If PHAST were to be adopted widely, the modest monthly subscription cost of $40-160 (4,000,000-16,000,000 translated characters per month) would be one of the only recurring expenses to using the system. It is most likely that a central authority, such as the state health department, would act as a host institution for PHAST and pay the subscription cost; it is also possible that member departments would share the cost among themselves. In either case, the monthly subscription
fee is at the low end of the cost range for a single translation by traditional methods ($130-1220). Its breadth of supported languages, low usage costs, and translation quality are the main reasons we used Microsoft Translator API instead of the Google Translate API, though a system like PHAST could be built using either option.

A single source document can be added to the system and translated into multiple languages at once using the upload page. At present, PHAST only supports unformatted .txt and .docx files, because it uses PHPDOCX to convert documents into text format that is interpretable by Microsoft Translator. The upload page’s interface contains text entry fields for the document’s title, topic, intended audience, desired reading level, and other notes. Below these fields, the user selects the file to upload and checks the appropriate boxes next to the MT output languages available. The translated versions are then added to the pool of documents that are ready for post-editing (see Figure 1) and are tracked through the post-editing process. PHAST sends an email to users who have language expertise in the target language, notifying them that a new document has been added and is ready for post-editing.

Users can start post-editing by “claiming” the document, which prevents others from post-editing it simultaneously (see [1] in Figure 2), protecting their work and allowing for version control. Any user in the system can claim a document, allowing staff from one health department to add a document and bilingual staff from another to claim and post-edit the document. Because bilingual public health staff often have so many competing demands and perform translation work between other tasks, it may take them several post-editing sessions to complete the process, or they may need to hand the document off to someone else to finish the post-editing. They may also need someone else to perform a quality check of their work. By clicking the “unclaim” button, other users can access the document to revise or finish the post-editing (2).

Once a user initiates the post-editing process (see Figure 2), PHAST automatically separates the documents into segments based on delimiters (line breaks) in the source document (3). For each segment, the source text (4) is shown above its corresponding MT, which is now available for editing. When the post-editor has finished editing a segment, the user saves it (5), prompting the system to highlight the segment in green. During our interviews, we found that public health workflow requires intermittent post-editing and that staff require a method for easily saving their progress in order to resume post-editing at a later time. The “postedited” bar at the top of the screen tracks a document’s progress through the translation process (6), while the green highlighted sections remind post-editors which parts they have already worked on. The post-editing interface includes a discussion board where users can post comments about the document for reference (7). Our user interviews indicated that staff requiring a document translation may need to communicate particular instructions to the post-editors. For example, they could resolve disagreements about how to translate specific parts of the document.

Figure 2: Post-editing interface
Evaluation

We have studied the PHAST tool’s functionality and usability with public health department staff in Washington State. Prior to usability testing, we performed unit testing, whereby we tested the functionality of each feature in isolation; scenario testing, whereby we identified and performed a set of scenarios and verified that the system produced the desired output; and stress testing, whereby multiple users logged into the system at once and performed tasks with the desired outcome.

Before presenting our prototype to target users, we performed early usability testing with three bilingual volunteers who, like health department staff, were not professional translators. Appropriate changes were made and, when PHAST was determined to be ready for our target users, we presented it to a Washington Department of Health focus group comprising various health education and translation staff. Participants validated the translation workflow around which PHAST was designed, watched a demonstration of the system, and tested the main functions of the PHAST website. We noted critical incidences and user preferences to inform changes to PHAST. Once we made this series of initial changes, we performed formal user testing. Staff from two public health agencies in Washington were given scenarios based on the workflow findings and asked to perform a series of tasks that reflect the main PHAST functions and the typical translation workflow. Our one-on-one, in-person user tests with health department translation staff took place at their offices, on the current live PHAST website, phastsystem.org.

We would like to emphasize that fully testing the collaborative aspects of PHAST would require a complete deployment to several health departments. This would require a push from a state or federal authority, which was beyond the scope of this research project. Instead, we used a novel strategy for groupware usability testing to simulate two components of statewide deployment: the test scenarios and an analysis framework focused on evaluating collaboration. The test scenarios took users through the collaborative aspects of PHAST and encouraged them to think about performing each of the tasks within the context of inter-agency cooperation (e.g. “You want to leave the document for someone else to continue editing. How would you do that?”). As part of the analysis, we used Gutwin et al.’s mechanics of collaboration framework to evaluate how well PHAST in its current form supports collaboration.

In total, we completed the user tests with 10 staff members from two mid-sized health departments. All participants were involved in producing translated health materials in some capacity, but their job titles varied and included communications, education, nursing, and design roles. Their experience with translation work and technical expertise also varied widely; seven of the 10 had tried statistical MT systems, primarily Google Translate. Participants sat at laptops in their departments’ respective conference rooms to perform tasks and were video recorded to capture audio and actions on the laptop screen. In addition to the moderator, there was a note taker present for each session. Participants were asked to think aloud and were asked probe questions when appropriate.

For this reason, we did not report the time spent on each task. However, it is worth noting that the time required to complete all tasks ranged from around 7 to 28 minutes, depending on whether the users encountered difficulties and on the extent of the think-aloud commentary and probes. Data were collected on task success/failure; error count; error severity on a scale of 1-5, with 1 indicating completion with no problems and 5 indicating failure; as expected/not as expected statements; and perceived satisfaction, with 1 indicating completing the task with no problems and statements that it was exactly as expected or better and 5 indicating failure, with complaints.

The tasks were situated within scenarios based on our translation workflow findings and were read by the moderator. The scenarios emphasized the collaborative aspects of PHAST, in order to simulate realistic interactions across health departments that would likely occur if it were deployed widely. In order, the tasks were: login, identify who’s working on a document, identify the progress bar showing how complete the post-editing is, upload a document and fill in the metadata fields associated with it, claim a document, post-edit three lines of the claimed document and save them, identify the lines that have been edited and saved, leave a note for another post-editor in the discussion box, unclaim the document, and download a completed document. After each task, the moderator asked, “Is that what you would expect?”

Data Analysis

We analyzed the data by code in Excel (e.g., error count per task and participant) and with the code co-occurrence table feature in Atlas.ti, in order to find correlations between codes (e.g. claim a document and as expected/not as
expected). Three members of the study team used Atlas.ti to blindly code the videos. We used a closed codebook containing three main types of codes: codes for each task; codes for the aforementioned metrics that point to effectiveness, efficiency and satisfaction with PHAST; and codes for each of the mechanics of collaboration (explicit communication, consequential communication, coordination of action, planning, monitoring, assistance, and protection). Though some mechanics of collaboration were present across several tasks, such as coordination of action, we found planning to be too high-level for the purposes of evaluating PHAST, and therefore excluded it. We also coded suggestions and comments on heuristics. Open coding was encouraged for interesting quotes and additional themes, though it did not ultimately produce any useful findings. We merged the blindly coded Atlas.ti hermeneutic unit files and cleaned the merged version for overlaps and redundant coding.

Results

Participants voiced satisfaction with PHAST by indicating that each task occurred as they expected in the vast majority of instances, and four participants explicitly said they would like to use such a system. One of the less technically proficient participants, P5, told the moderator: “I just think it would be an incredible tool for us. And I like the fact that you have whatever document you have and that people can get on and they can claim it and they can do their thing and then they can unclaim it and someone else can go look at that. That’s very easy.”

About 66% of errors were concentrated in four of the ten tasks: uploading a document, claiming a document for post-editing, post-editing and saving three lines, and unclaiming a document. We also found errors in downloading a completed translation, but all were due to the participant not understanding the meaning of the term “download.” A summary of the errors by task and participant is in Table 1 below.

<table>
<thead>
<tr>
<th>Task</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
<th>Total by task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Login</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Identify who’s working on a document</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>How complete is post-editing</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Upload a document</td>
<td>2</td>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Claim</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Post-edit</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Identified lines edited and saved</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Leave a note</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Unclaim</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Download</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td><strong>Total by participant</strong></td>
<td>1</td>
<td>10</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 1: Errors by task and by participant

The “uploading a new document” task had a high number of errors and two failures out of ten participant attempts. The errors were concentrated amongst the seemingly less technically proficient participants; four of the ten did not understand what “upload” means and needed prompts or further explanation to add a document to the system. Just over half of all errors were from two participants.

Claiming a document had the second-highest number of errors. While all ten participants completed the task successfully, five needed prompting. A majority voiced confusion over the need to claim the document before post-editing. We believe that this task has a high degree of trainability and that users would not have issues in subsequent attempts. Much of the confusion stemmed from the terminology of claim/unclaim; participant 10 said: “I’m not sure

497
that claimed is the best word. Probably like a library checked in and checked out... We use SharePoint here and I think that’s also a common terminology with SharePoint is checking out and checking in a document.”

The vast majority of failures stemmed from two of the users. Post-editing and saving a line had three failures, all of which stemmed from the participant not saving an edited line before moving on to the next. Identifying edited and saved lines also had three failures, one of which was due to not recognizing that the saved lines changed color. We believe this has a high degree of learnability and would not be an issue in repeated use. This task also evoked many suggestions for a track change feature or other ways to see what other post-editors have changed in the document. Unclaiming a document had two failures due to the users logging out or closing the window instead of clicking the unclaim button. This could cause critical bottlenecks to a translation’s progress through the system and will require design changes. A summary of the failures by task and by participant is in Table 2 below.

<table>
<thead>
<tr>
<th>Task</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
<th>Total by task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Login</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Identify who’s working on a document</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>How complete is post-editing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Upload a document</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Claim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Post-edit</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Identified lines edited and saved</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Leave a note</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Unclaim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Download</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td><strong>Total by participant</strong></td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>13</td>
</tr>
</tbody>
</table>

*Table 2: Failures by task and by participant*

To evaluate whether PHAST allows users to work collaboratively to complete common translation tasks, we plotted each of the mechanics of collaboration codes against the effectiveness (success/failure), efficiency (error count and severity) and satisfaction (as expected/not as expected and satisfaction 1-5) codes in Atlas.ti’s code co-occurrence table, in a novel groupware prototype evaluation method.\(^{27}\) The table gives a correlation coefficient that shows the degree to which two codes overlap. We used the table’s correlation coefficients qualitatively, as a way of finding themes by drawing attention to the extreme values, rather than as correlation coefficients per se. The login and download tasks were not considered collaborative enough for inclusion in this part of the analysis, so only eight of the ten tasks were analyzed. We also excluded the “planning” mechanic code, after deciding it was too high level for PHAST.

We found that PHAST adequately supports explicit communication, as embodied in the document upload and leave a note tasks. Effectiveness, efficiency and satisfaction were all deemed adequate because the positive codes (success, error 1-2, as expected, satisfaction 1-2) were the most strongly associated with the explicit communication code. The two upload failures were related to technical proficiency, rather than to any major design flaws, while all participants were successful in leaving a note for a fellow post-editor in the discussion box. Both tasks had high co-occurrence with positive satisfaction codes. Aside from leaving a note, all tasks included in this analysis involved consequential communication; the results for this code were thus somewhat mixed. The code co-occurrence table showed consequential communication to have a great degree of overlap with both success and failure codes. While there were 30+ errors in the tasks considered consequential communication, two participants accounted for more than half of them, so efficiency was also mixed. Satisfaction was high: the “as expected” code occurred with a much
greater frequency than “not as expected” and satisfaction scores 1 and 2 were the most frequently associated with consequential communication.

Similarly, the codes “coordination of action” and “monitoring” spanned all tasks included in this analysis. This is mostly due to the fact that the entire system and corresponding scenarios follow the translation workflow model, and these mechanics are embedded to some degree in every task. We found need for improvement in both, which is in line with our overall individual user experience findings. In terms of effectiveness, there were failures in seven of the eight tasks, with two or more each in uploading, post-editing, identifying edited and saved lines, and unclaiming. Coordination of action showed only moderate overlap with both success and failure in the code co-occurrence table. The success code was most strongly associated with monitoring, and failures had the third-strongest association. With regard to efficiency, there were 30+ errors, over the ten participants, in both coordination of action and monitoring, but the severity was relatively low and, as mentioned above, two participants accounted for more than half the errors. Satisfaction, on the other hand, was relatively high, with “as expected” and satisfaction scores of 3 and above being the most commonly associated with the coordination of action and monitoring codes.

PHAST has adequate support for assistance, which was practiced with the task of leaving a note in the discussion box. There were no failures (effectiveness) and no errors (efficiency), but participants made several suggestions for track changes or line-by-line commenting. Satisfaction was mixed, with “not as expected” co-occurring slightly more frequently than “as expected” and satisfaction scores mostly around 3.

Protection was part of all tasks except uploading a new document, and support for this code was mixed. There were seven failures, so effectiveness for tasks considered protection could be improved. Efficiency was mixed, with 24 errors but low severity ratings. Satisfaction was moderately high, with a greater frequency of “as expected” than “not as expected,” and a majority of satisfaction scores clustered around 1-3.

Discussion and implications for design

The usability testing findings indicate that PHAST could be successfully integrated in public health practice and that public health staff are receptive to using such a system. Still, additional changes should be made prior to any full deployment of the system. The vast majority of changes are relatively minor adjustments to the user interface, e.g., revising the use of claim/unclaim and prompting users to unclaim a document before logging out or closing the browser if they intend to stop post-editing at that time. These types of changes could increase PHAST’s usability and consequently improve its support for collaboration. Even with a design process that emphasized the needs of lay users, the findings show that the system could still be difficult for less technically proficient users in their first encounter. While we accounted for this with the step-by-step translation flow, highly learnable features, and design updates such as changing the upload page name from “Upload a document” to “Add new document,” there are limitations to how much the design can accommodate users unfamiliar with common terms such as “download.” Additional training may be required for these users.

In reviewing the findings and considering them in the context of our past and ongoing interactions with public health translation staff, we believe developing a way to better track changes or post-edited versions may be critical to the success of a full deployment. In the earlier workflow studies and in the post-editing tasks, public health employees voiced concern about quality assurance. This concern surfaced again during the user testing; one participant said, “My concern is that, will it save it as this is the change that I’ve made, but if somebody else didn’t want that change, the other version was still there... If I made a change and somebody else came back and said ‘Well what happened to my version?’ You know, I just think there needs to be some documentation there.” While quality assurance is also a concern with traditional translation vendors, our prior studies showed that health departments formed relationships with trusted vendors. A similar relationship could be built into PHAST with post-editor ratings by users. Post-editors would then have a greater incentive to produce quality translations and departments uploading documents would have more control over quality assurance. This addition would require careful consideration so as not to produce excessive demand for certain post-editors while discouraging others or damaging morale.

These concerns over quality assurance and respect for bilingual employees’ workload are compounded by the funding structure of local public health in the US, which generally occurs at the county or city government level and thus funding authority is highly decentralized. It would likely take a boost from a central authority to fully
implement and maintain PHAST in Washington State. To this end, we have built an ongoing partnership with the
Washington Department of Health, which has voiced interest in hosting and driving the adoption of PHAST.

Conclusion, limitations and future work

We have presented a collaborative MT system that can significantly reduce the time and cost of producing
multilingual health materials and was deemed usable by current public health staff. The system’s collaborative
aspects provide a new way to share limited translation and post-editing resources via a groupware system. We have
relied on user-centered design methods to iteratively develop PHAST and to better ensure that it meets the needs of
lay users and their work context; and usability testing results indicate that additional minor training would be
required only for the least technically proficient users. We identified design requirements that should be
implemented before attempting a full deployment, including additional quality assurance features. Offline support
policies, such as a partnership with a central authority that can help drive adoption, would also be necessary for full
implementation.

PHAST has been developed based on user studies with personnel from health departments in WA State. Our
findings may not reflect the practices of health departments in other states. A full evaluation of PHAST’s support for
collaborative work was limited by the fact that it was not yet fully implemented across health departments.
However, this provided an opportunity to simulate and then evaluate the system with the mechanics of collaboration
framework, which we believe to be a useful contribution to our study and to the public health informatics
community.

Statistical MT combined with post-editing has already been shown to reduce errors after machine translation is
performed. Currently, PHAST uses a generic MT engine (Microsoft Translator, (http://www.microsoft.com/en-us/translator/)). The Microsoft Translator Hub (http://hub.microsofttranslator.com/) allows users to train their own
customized models for a particular domain or style of text, provided they can supply a sufficiently large training set
(a minimum of 2000 sentences, though larger sets will yield better results). We have run initial experiments with
customized models produced by the Translator Hub, and we have found that the translation performance (as
measured by automatic evaluation procedures) is improved over generic models. A more widespread and regular
deployment of the PHAST system would enable the collection of a large number of examples of translated
materials, which in the future can be used to generate more accurate automated translation models. Thus, one of the
long-term goals of this work is to collect a sufficient number of quality translated documents to produce customized
translation models. Currently, the PHAST system does provide functionality for collecting documents it has
processed in a parallel text format, but the procedure for building new models has not been automated or integrated
into the PHAST system. This growing collection would allow future development of dynamic, customized MT in
the public health domain and could be replicable for other domain-specific MT systems.

Acknowledgements

This study was funded by grant #1R01LM010811-01 from the National Library of Medicine (NLM). Its content is
the sole responsibility of the authors and does not necessarily represent the view of the NLM. We would like to
thank the WA health department staff members who volunteered their time during various stages of this project.

References

1. Pandya C, McHugh M, Batalova J. Limited English Proficient Individuals in the United States: Number, Share,
Growth, and Linguistic Diversity. LEP Data Brief. Migration Policy Institute, 2011.
2. Peña-Purcell N. Hispanics’ use of Internet health information: an exploratory study. Journal of the Medical
4. Cheng EM, Chen A, Cunningham W. Primary language and receipt of recommended health care among
625-642.


Supervised learning is the dominant approach to automatic electronic health records-based phenotyping, but it is expensive due to the cost of manual chart review. Semi-supervised learning takes advantage of both scarce labeled and plentiful unlabeled data. In this work, we study a family of semi-supervised learning algorithms based on Expectation Maximization (EM) in the context of several phenotyping tasks. We first experiment with the basic EM algorithm. When the modeling assumptions are violated, basic EM leads to inaccurate parameter estimation. Augmented EM attenuates this shortcoming by introducing a weighting factor that downweights the unlabeled data. Cross-validation does not always lead to the best setting of the weighting factor and other heuristic methods may be preferred. We show that accurate phenotyping models can be trained with only a few hundred labeled (and a large number of unlabeled) examples, potentially providing substantial savings in the amount of the required manual chart review.

Introduction

Mining massive databases of electronic health records for patients who satisfy a set of predefined criteria is known in medical informatics as phenotyping. Phenotyping has numerous use cases: clinical trial recruitment, outcome prediction, survival analysis, and other kinds of retrospective studies [1]. Several translational science initiatives have made identifying patient phenotype cohorts from the Electronic Health Records (EHR) their primary focus. Among them are Electronic Medical Records and Genomics (eMERGE) [2], Pharmacogenomics Network (PGRN) [3], and Informatics for Integrating Biology and the Bedside (i2b2) [4]. Within each of these initiatives, disease-specific phenotyping algorithms are developed and run against repositories containing millions of patient records. The identified patient cohorts are subsequently linked to biobanks for genetic analysis.

Supervised machine learning is currently the predominant paradigm in phenotyping [1]. Unfortunately, supervised learning can be very costly due to the expenses associated with the manual chart review. A separate machine learning model is developed for each phenotype, typically requiring hundreds of manually labeled examples. For instance, about 600 hundred patient records had to be reviewed manually for each disease within the i2b2 initiative. At the same time, thousands more were left unlabeled. Investigating cheaper alternatives to supervised learning that take advantage of the bountiful unlabeled patient records thus holds great promise.

Semi-supervised learning is a class of methods that is concerned with incorporating unlabeled data within machine learning models along with the data that have labels. In contrast with the general domain, semi-supervised learning has not received sufficient attention in the clinical NLP community. In this paper, we report on our experiments with several versions of the expectation maximization (EM) algorithm [5], a simple, yet powerful approach for semi-supervised learning. We experiment with four phenotyping datasets developed within the i2b2 initiative and analyze the interaction between the key aspects of semi-supervised learning that determine its success: the amount of labeled and unlabeled data and the relative weight of the unlabeled data. For our experiments, we adapt and implement several flavors of the EM algorithm [6,7] for modeling the i2b2 phenotypes. We first experiment with the basic EM algorithm. While the outcome of this experiment is promising, it becomes clear that unlabeled data often overwhelms the model, resulting in suboptimal parameter estimates. This shortcoming is ameliorated by introducing the weighting factor, allowing the unlabeled data to help characterize the parameter space without overwhelming the relatively small number of high quality gold labels. We investigate several alternatives for setting the value of the weighting factor including cross-validation and a simple heuristic. We discover that cross-validation is not always the optimal approach with the competing methods often leading to better models.

Our results indicate that overall unlabeled data can be highly beneficial to the models. We show that high-quality models can be trained with only a few hundred labeled examples when combined with unlabeled instances, potentially providing large savings in the amount of the required manual chart review. We also discuss several
practical lessons that can be drawn from our experiments. Our work is a bid to bring semi-supervised learning to the 
attention of the community by detailing a simple yet versatile baseline method.

The only study we are aware of that touches the subject of semi-supervised learning in the context of clinical NLP is 
the work by Garla et al. [8], who look at an application of Laplacian SVMs for detecting the presence of malignant 
liver lesions. While this work presents an encouraging result, it has several shortcomings that make it difficult to 
predict how well their findings would generalize to other tasks. First, the liver lesion model is based on complex 
rule-based features, whereas our model utilizes a simple bag of Unified Medical Language System (UMLS) [9] 
features that are successful in phenotyping [1,10–12]. Second, Garla et al. do not examine the effects of varying the 
amount of labeled data, which in our work leads to valuable insights about the practical aspects of semi-supervised 
learning. Finally, Garla et al. study a single dataset, employing a model with a large number of tunable parameters. 
In contrast, our work involves multiple datasets, while utilizing a simple and scalable model with a single tunable 
parameter (the weighting factor), the effects of which we carefully study. While our experiments with multiple 
datasets paint a more complex picture concerning the effects of unlabeled data, our conclusions are likely more 
generalizable and complete.

This paper is organized as follows. We first discuss our models and the EM algorithm. We then outline an 
experiment providing the intuition why EM works. Next, we examine the effects of varying the amount of labeled 
and unlabeled data and different methods for setting the weighting factor. Finally, we discuss our findings and draw practical lessons from our results.

Methods

Data representation

We perform our study in the setting where the unit of phenotype classification is the complete patient chart. We 
represent each chart as a set of UMLS concept unique identifiers (CUIs) which we extract from the patient’s records 
using Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES; ctakes.apache.org) [13]. CUIs 
are commonly used as representations in clinical NLP [12,14,15] as they help to abstract from lexical variability of 
medical terminology and capture the clinically relevant terms. Each CUI can be either asserted or negated, as 
determined by the cTAKES negation module. Some CUI examples employed in one of our models can be seen in 
Table 3.

Although cTAKES is capable of extracting most CUIs that exist in the UMLS, we only include the CUIs listed in 
phenotype-specific dictionaries. The dictionaries are created manually by i2b2 domain experts and define the 
relevant terms for each phenotype. Each phenotype-specific dictionary has \( D \) entries: \( CUI_1, \ldots, CUI_D \). We model a 
patient \( i \) as a \( D \)-dimensional vector of CUIs \( \mathbf{x}_i \) in which an element \( x_{ik} \) indicates how many times \( CUI_k \) from 
the dictionary was seen in the patient’s chart. Wu et al. [16] refer to this as “sum aggregation” and found it to perform 
well across performance metrics in a task requiring aggregating representations of patient charts.

Model

In the setting where all data have labels, model parameter estimation is fairly straightforward: for example, such 
methods as maximum likelihood or maximum a posteriori estimation can be utilized. When some of the labels are 
not known, parameter estimation is harder, although methods for obtaining parameter estimates exist. One such 
method is the Expectation Maximization (EM) algorithm [5]. The EM algorithm is an iterative procedure that begins 
by estimating model parameters from labeled data only. The model is then used to assign probabilistically-weighted 
class labels to the unlabeled instances. The model parameters are subsequently re-estimated from all (labeled and 
unlabeled) data and the procedure repeats until the parameter estimates stabilize. The algorithm is outlined in Table 
1.
• **Input**: labeled and unlabeled data
• Estimate model parameters from labeled data only (Equations 3 and 4)
• Loop until convergence
  o **E-step**: use current model parameters to compute the class distributions for patients with no labels (Equation 2)
  o **M-step**: re-estimate model parameters from both labeled and unlabeled examples (Equations 3 and 4)
• **Output**: model parameter estimates

<table>
<thead>
<tr>
<th>Table 1. Basic EM Algorithm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The probabilistic framework we utilize in conjunction with the EM algorithm is based on a multinomial Naïve Bayes classifier, an approach frequently used in the context of text classification [17]. Phenotyping is cast as a binary patient classification task where a patient ( x_i ) has a class label ( c \in {\text{case, noncase}} ). The total number of patients in the training set is ( T = T_{\text{labeled}} + T_{\text{unlabeled}} ), where ( T_{\text{labeled}} ) and ( T_{\text{unlabeled}} ) are the number of labeled and unlabeled examples respectively.</td>
</tr>
<tr>
<td>The posterior probability distribution over the labels for a patient ( x_i ) is given by Bayes rule:</td>
</tr>
</tbody>
</table>
| \[
    p(c | x_i) = \frac{p(c)p(x_i | c)}{p(x_i)} \quad (1)
| |
| Incorporating the standard Naïve Bayes assumption, this equation can be factorized to include class-specific CUI distributions \( p(\text{cui}_k | c) \) and the prior class probabilities \( p(c) \) as model parameters: |
| \[
    p(c | x_i) = \frac{p(c) \prod_{k=1}^{D} p(\text{cui}_k | c)^{x_{ik}}}{\sum_{c' \in \{\text{case, noncase}\}} p(c') \prod_{k=1}^{D} p(\text{cui}_k | c')^{x_{ik}}} \quad (2)
| |
| The parameters of this model can be computed using maximum a posteriori estimation: |
| \[
    p(\text{cui}_k | c) = \frac{1 + \sum_{i=1}^{T} x_{ik} p(c | x_i)}{D + \sum_{d=1}^{D} \sum_{i=1}^{T} x_{id} p(c | x_i)} \quad (3)
| \[
    p(c) = \frac{1 + \sum_{i=1}^{T} p(c | x_i)}{2 + T} \quad (4)
| |
| Notice that the generative model that gives rise to these equations assumes that the patients are generated from a mixture of two components that correspond to the two classes (case and noncase). These assumptions are likely to be violated in practice. The patient pool used for selecting the data to be labeled could potentially contain patients with other medical conditions whose class-specific CUI distributions could not be captured well by model parameters \( p(\text{cui}_k | c) \). It could also be that the “natural” (highest probability) clustering of the data does not fit well into the case/noncase class boundaries we adopted for the phenotyping task. Depending on the extent to which the modeling assumptions are violated, the unlabeled data could prove to be detrimental to the performance of the model. In order to alleviate the negative impact of the violated assumptions, Nigam et al. [7] augment their model with a parameter \( \lambda \) that determines the contribution of the unlabeled data. We investigate several strategies for setting this parameter and their effect on the quality of the ensuing model. |
| The augmented version of EM is written as follows. We first define \( \lambda_i \) to be the weighting factor \( \lambda \) for the unlabeled examples and 1 for the labeled examples. We then rewrite the model parameters as follows: |
| \[
    p(\text{cui}_k | c) = \frac{1 + \sum_{i=1}^{T} \lambda_i x_{ik} p(c | x_i)}{D + \sum_{d=1}^{D} \sum_{i=1}^{T} \lambda_i x_{id} p(c | x_i)} \quad (5)
This augmented EM algorithm is identical to basic EM in Table 1 except Equations 3 and 4 are swapped for Equations 5 and 6 respectively. The weight of the unlabeled data \( \lambda \) can be set via n-fold cross-validation, splitting the labeled data into a training and test sets and attempting different values of lambda. In addition to cross-validation, we introduce a heuristic that sets the weighting factor using a simple equation:

\[
\lambda = \frac{1}{T_{labeled}} 
\]

The rationale for this heuristic is the following: the model should be able to obtain accurate parameter estimates from the labeled data alone, when a large amount of it is available. Thus the weighting factor should decrease as the amount of labeled data increases.

Datasets
We utilize four datasets all of which were created within the i2b2 initiative [12,15,18,19]. We show various important characteristics of our datasets in Table 2.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Cohort size (patients)</th>
<th>Manually labeled Instances (patients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ulcerative Colitis (UC)</td>
<td>33,465</td>
<td>600</td>
</tr>
<tr>
<td>Crohn’s Disease (CD)</td>
<td>33,465</td>
<td>600</td>
</tr>
<tr>
<td>Multiple Sclerosis (MS)</td>
<td>172,447</td>
<td>595</td>
</tr>
<tr>
<td>Type II Diabetes (T2D)</td>
<td>198,002</td>
<td>596</td>
</tr>
</tbody>
</table>

Table 2. Phenotyping datasets used in the experiments.

Domain experts defined the ICD-9 codes relevant for each phenotype. These were then used to create the initial cohort from more than 6 million patient EHRs of the Partners Healthcare System. From that initial cohort, about 600 patients were randomly chosen for manual labeling. Each patient chart was reviewed by a domain expert and labeled at the patient level.

Obviously, only a small portion of the initial cohort could be labeled manually, leaving a large number of unlabeled examples for our semi-supervised learning experiments. For each experiment, we sampled randomly 500, 1000, and 3000 patients from the portion of the cohort that was never labeled.

Experimental Setup
The main objective of our evaluation is to determine whether unlabeled data can improve the classification accuracy of supervised models. The most natural baseline in these circumstances is, therefore, a supervised learning baseline in which the models are trained using labeled data only. During evaluation, we examine the behavior of semi-supervised learning with respect to the supervised baseline. This type of evaluation is best conducted via an analysis of learning curves as this allows us to compare the performance of supervised and semi-supervised models at different sizes of the training set.

To generate smoother curves, we utilize 10-fold cross-validation. In a typical experiment, within each fold we allocate a held-out test set and a pool of examples from which the labeled data is sampled. To produce a single point on a learning curve, we average the performance on the held-out test sets across all folds. To produce a point on the baseline curve, we train a Naïve Bayes classifier using labeled data only and evaluate its performance on the held-out test set. To produce a point on the semi-supervised learning curve, we add the unlabeled data, run the EM algorithm for 25 iterations\(^1\), and use the resulting model to classify the test set. The resulting curves can be

\[^1\) In our preliminary experiments, EM typically converged after 15-20 iterations.
compared visually or numerically in terms of the area under the curve (AUC) or increase in accuracy. For each phenotype, we generate the supervised baseline and semi-supervised learning curves for 500, 1000, and 3000 unlabeled examples.

We first experiment with the basic EM algorithm, which could also be viewed as a version of the augmented EM in which the weight of the labeled and unlabeled examples is the same. To illustrate the inner working of the EM algorithm, we examine the feature weights across EM iterations. Next, we experiment with augmented EM, down-weighting the unlabeled examples. Finally we experiment with setting the weight parameter using our heuristic technique and 10-fold cross-validation. During cross-validation, the held-out test set is, of course, not used; instead a validation set is allocated within each fold and each of the following lambda values is evaluated: 0, 0.05, 0.25, 0.50, 0.75, 1.0.

Due to the limit on the number of figures, we are not able to provide individual learning curves for the four phenotypes for all experimental conditions. Instead, we summarize each condition by plotting the curves that are averaged across all four phenotypes. These plots are supplemented by individual phenotyping plots for several important experimental conditions. All experiments reported in this paper are based on our own implementation of the EM algorithm and the supervised baseline.

Results
The averaged learning curves obtained in the experiment with basic EM algorithm are shown in Figure 1 (upper left).

![Learning Curves](image)

Figure 1. Average Learning Curves

The individual learning curves for each of the four phenotypes are in Figure 2. To take a peek “under the hood” of the EM algorithm, we train a Crohn’s Disease model using 10 labeled examples. The top ten features selected using log-likelihood ratio are shown in the first two columns of Table 3. The top ten features selected using log-likelihood ratio after running EM for 25 iterations with 3000 unlabeled examples are in the last two columns of Table 3.
Labeled data only | Labeled and unlabeled data
---|---
| CUI | Description | CUI | Description |
c0006826 | Neoplasms malignant | -c1171255 | Humira |
c0032952 | Prednisone | c2343521 | Cimzia |
c0030193 | Pain | c1172734 | Natalizumab |
c0001418 | Adenocarcinoma | c0001418 | Adenocarcinoma |
c0009324 | Ulcerative colitis | c1171255 | Humira |
c0678172 | Asacol | c1872109 | Certolizumab pegol |
-00007097 | Carcinoma | c2343521 | Cimzia |
-1292819 | Resection | c0162529 | Ischemic colitis |
c0007097 | Carcinoma | -c0086492 | J pouch |
c0009410 | Colostomy | c0678171 | Pentasa |

Table 3. Top ten features before and after execution of EM for Crohn’s Disease. The ‘-’ preceding the CUI indicates negation.

The results of running the augmented EM algorithm with $\lambda$ set to 0.05, 0.20, and 0.50, are in the bottom row of Figure 1. We also provide the individual learning curves for $\lambda=0.05$ in Figure 3. Finally, the results of using our $\lambda$ selection heuristic and 10-fold cross validation are shown in Figure 1 (top middle and top right); the individual phenotype learning curves for these conditions are also available in Figures 4 and 5 respectively.

One way to summarize these plots numerically is to examine the difference in the area under the curve (AUC) between the learning curve of a semi-supervised approach and the supervised baseline. Another is to compute the average improvement of a semi-supervised learning curve over the learning curve for the supervised baseline. We show both in Table 4 with the two metrics reported across phenotypes, number of unlabeled examples, and lambda selection criteria. The differences in AUC and the average improvement were computed across all training set sizes.

<table>
<thead>
<tr>
<th></th>
<th>$\lambda = 1.00$</th>
<th>$\lambda = 0.05$</th>
<th>$\lambda$ selection heuristic</th>
<th>$\lambda$ cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>3.50</td>
<td>8.61</td>
<td>7.99</td>
<td>1.27</td>
</tr>
<tr>
<td>UC</td>
<td>2.22</td>
<td>-1.56</td>
<td>-5.59</td>
<td>2.79</td>
</tr>
<tr>
<td>MS</td>
<td>6.92</td>
<td>-1.43</td>
<td>-14.99</td>
<td>6.28</td>
</tr>
<tr>
<td>T2D</td>
<td>4.12</td>
<td>6.91</td>
<td>5.86</td>
<td>1.61</td>
</tr>
<tr>
<td>Average</td>
<td>1.88</td>
<td>4.14</td>
<td>2.49</td>
<td>4.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\lambda = 1.00$</th>
<th>$\lambda = 0.05$</th>
<th>$\lambda$ selection heuristic</th>
<th>$\lambda$ cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>0.91</td>
<td>2.31</td>
<td>2.19</td>
<td>0.32</td>
</tr>
<tr>
<td>UC</td>
<td>0.60</td>
<td>-0.47</td>
<td>-1.57</td>
<td>0.79</td>
</tr>
<tr>
<td>MS</td>
<td>1.98</td>
<td>-0.22</td>
<td>-3.79</td>
<td>1.84</td>
</tr>
<tr>
<td>T2D</td>
<td>1.05</td>
<td>1.76</td>
<td>1.56</td>
<td>0.39</td>
</tr>
<tr>
<td>Average</td>
<td>0.53</td>
<td>1.14</td>
<td>0.70</td>
<td>1.15</td>
</tr>
</tbody>
</table>
Table 4. Difference in AUC and average improvement of semi-supervised learning over the supervised baseline.

Discussion

Incorporating unlabeled data into the model using basic EM algorithm shows very promising results, as evidenced by Figure 1 (top left). For Crohn’s Disease dataset (Figure 2, top left), the fully supervised model trained on 10 labeled examples is only 74% accurate. With 3000 unlabeled examples added, the model receives a large performance boost: the accuracy is now 85%. It takes more than 300 more labeled examples for the supervised model to achieve the same level of performance. Clearly semi-supervised learning has a great potential, offering a large reduction in the amount of manual annotation that is required to train an accurate model.

Why does EM algorithm work? The top ten features (Table 3, left) learned by the supervised model are simply too general to reliably identify a Crohn’s Disease patient: notice the “Pain” and “Prednisone” features (“Prednisone” is a drug used to treat many inflammatory diseases). On the other hand, adding unlabeled data (Table 3, right) leads to a feature set that is much more specific to Crohn’s Disease: “Pain” no longer appears on the list; Humira, Cimzia, Natalizumab, Certolizumab pegol, and Pentasa are all medications that are used to treat this condition. With few labeled examples, there is apparently not enough signal to capture the high number of drug names. However, by adding data that cluster with the positive labeled examples, the drug name features are found to be highly discriminative. Thus, unlabeled data helps to obtain better model parameter estimates.

However, the behavior of the remaining three phenotypes in Figure 2 paints a more complex picture. To remind the reader, Figure 2 presents the performance of the models where labeled and unlabeled instances are both given the same weight. While injecting a small amount of unlabeled data (500 unlabeled examples) improves the model performance, adding a larger amount (1000 and 3000 examples) is detrimental to classification accuracy for two out of four datasets we tried (Ulcerative Colitis and Multiple Sclerosis). This effect of unlabeled data is likely caused by violated model assumptions. When a large number of unlabeled examples is present, model parameter estimation is strongly influenced by the counts from the unlabeled data. The model essentially performs unsupervised clustering, using the labeled examples mostly to assign cluster memberships. When the most probable two-class clustering of the data does not match the annotated class boundaries, the unlabeled data can skew parameter estimation, potentially even leading to worse estimates than with labeled data only.
Introducing the $\lambda$ parameter that downweights the unlabeled data leads to more accurate model parameter estimation. As we decrease $\lambda$, injecting additional unlabeled examples should more thoroughly describe the space we are classifying in without swamping the highly valuable gold standard instances. This results in more consistent improvements over the supervised baseline. Observe that basic EM (Figure 1, top left) did not seem to take full advantage of unlabeled data: increasing the amount of unlabeled data from 500 to 3000 resulted in a reduction in accuracy. By lowering the value of $\lambda$ (Figure 1, bottom row), EM makes increasingly better use of unlabeled data, pushing the 3000 curve higher. At $\lambda = 0.05$ (Figure 1, bottom left; Figure 3) as the amount of unlabeled data is increased to 3000 examples, the semi-supervised models beat the supervised baseline for most training set sizes.

Applying our lambda selection heuristic results in a similar behavior (Figure 1, top center; Figure 4), although the overall gains over the supervised baseline are smaller both in terms of the AUC (2.49 vs. 4.14 for $\lambda = 0.05$) and the average improvement in accuracy (0.70 vs. 1.14 percent average improvement).

Using 10-fold cross-validation for lambda selection (Figure 1, top right; Figure 5) brings about overall gains that are comparable to using a small lambda, both in terms of AUC and the average improvement in accuracy. However, cross-validation does not appear to consistently pick the best lambda. For example, for the Crohn’s Disease dataset, it is possible for the semi-supervised models to do better if $\lambda = 1.00$ is selected for all amounts of unlabeled data we evaluated, in terms of both AUC and average accuracy improvements. For Multiple Sclerosis (see Table 4), cross-validation does not seem to take full advantage of the unlabeled data: the model that has access to 500 unlabeled examples outperforms the models that have access to more unlabeled data. At the same time for $\lambda = 0.05$ the best result is achieved by the model that has 3000 unlabeled examples; that model also outperforms all models selected via cross-validation in terms of both AUC (6.74) and average accuracy improvement (1.92). Surprisingly, even the simple lambda selection heuristic makes better use of unlabeled data than cross-validation: the models that utilize the heuristic, outperform the models where lambda was selected through cross-validation when 1000 and 3000 unlabeled examples are included.

Overall, whether lambda is computed using our simple heuristic, using 10-fold cross-validation, or simply set to some small value, the unlabeled data has the strongest positive effect on model performance when the amount of labeled data is relatively small. For $\lambda = 0.05$ (Figure 1, bottom left) the best semi-supervised model reaches its top performance at around 150 labeled examples. It takes more than twice as many labeled examples for the supervised model to reach the same level of performance. Eventually, as the number of labeled examples in the training set grows, the models are capable of finding good parameter estimates from the labeled data alone and the learning curves for the supervised and semi-supervised cases become very similar.

These findings are relevant to scaling up phenotype algorithm development for projects such as eMERGE, PGRN, and now BD2K. Domain expert time is always limited, and if a maximum performance can be achieved by

![Figure 4. Learning curves for $\lambda$ set by heuristic.](image1)

![Figure 5. Learning curves for $\lambda$ set by cross-validation.](image2)
combining a small seed of label instances (100-200) with unlabeled data, that would lead to gains in efficiency. If the domain expert time is plentiful, then large amounts of labeled data would be preferable as the completely supervised model will reach accurate parameter estimates.

**Conclusion**

Even though for some datasets the basic version of EM did not succeed in beating the supervised baseline for larger amounts of unlabeled data, it appears to be possible to overcome this shortcoming. Introducing the lambda parameter that downweights the unlabeled data leads to more consistent gains over the supervised baseline. Clearly, semi-supervised learning can be of high importance in practice. Are there any practical lessons we can learn from the experiments we presented?

To answer this question, let us first identify two properties of a model that makes use of unlabeled data that a practitioner of semi-supervised learning would find desirable. First, a semi-supervised model should not perform worse than the corresponding supervised baseline. Clearly, if adding unlabeled data is as likely to lead to performance deterioration as to improvement, it is too risky to use in practice. Second, the semi-supervised model should make a full use of unlabeled data that is available to it. Out of two models, we prefer the one that can “squeeze” more out of unlabeled data, resulting in larger gains over the supervised baseline.

In light of these properties, we can draw several lessons from our experiments. It appears that semi-supervised learning of the type we discussed in this work has a higher chance of success if augmented EM is used. A priori, the most obvious approach to setting its lambda parameter would be to use n-fold cross-validation. Empirically, simply setting lambda conservatively (i.e. to some small value) works a little better, possibly due to the sensitivity of semi-supervised learning to the amount of labeled data (n-fold cross-validation has to allocate some portion of the labeled data as a validation set for parameter evaluation). The simple lambda selection heuristic we introduced also has some desirable properties over n-fold cross-validation: it often results in more consistent improvements over the supervised baseline and for larger training set sizes it results in performance level that is no worse than the supervised baseline. At the same time, it is much more efficient, which may be important for rapid development of multiple phenotyping models. Finally, the heuristic does not require the implicit supervision of a human-generated list of acceptable weights as in cross-validation. If the size of the unlabeled data were increased by an order of magnitude, the cross-validation procedure may require manual adjustments to its inputs to find a suitable value.

For future work, we are planning to examine the behavior of semi-supervised learning under the scenario where an order of magnitude more unlabeled data is available to the models. We will also investigate the use of Bayesian inference methods for model parameter estimation such as Gibbs sampling.

**Acknowledgements**

The study was funded by U54LM008748 (i2b2), 1R01GM103859-01A1 (PGx), 1U24CA184407-01 (DeepPhe), and R01GM090187 (ShARe).

**References**


POETenceph - Automatic identification of clinical notes indicating encephalopathy using a realist ontology

Kristina M. Doing-Harris, PhD\textsuperscript{1,2}, Charlene R. Weir, PhD\textsuperscript{1,2}, Sean Igo, MS\textsuperscript{1}, Jianlin Shi, PhD\textsuperscript{1}, Yijun Shao, PhD\textsuperscript{1,2}, John F. Hurdle, MD, PhD\textsuperscript{1}
\textsuperscript{1}University of Utah, Salt Lake City, UT; \textsuperscript{2}VA Salt Lake City Health Care System, Salt Lake City, UT

Abstract

Identifying inpatients with encephalopathy is important. The disorder is prevalent, often missed, and puts patients at risk. We describe POETenceph, natural language processing pipeline, which ranks clinical notes on the extent to which they indicate the patient had encephalopathy. We use a realist ontology of the entities and relationships indicative of encephalopathy in clinical notes. POETenceph includes a passage rank algorithm, which takes identified disorders; matches them to the ontology; calculates the diffuseness, centrality, and length of the matched entry; adds the scores; and returns the ranked documents. We evaluate it against a corpus of clinical documents annotated for evidence of delirium. Higher POETenceph are associated with increasing numbers of reviewer annotations. Detailed examination found that 65% of the bottom scoring documents contained little or no evidence and 70% of the top contained good evidence. POETenceph can effectively rank clinical documents for their evidence of encephalopathy as characterized by delirium.

Introduction

POETenceph detects encephalopathy in inpatient notes from all clinical and ancillary domains. Encephalopathy is a broad term for brain disease, damage or malfunction\textsuperscript{1}. It is characterized by an altered mental state. Delirium is a sudden encephalopathy or state of confusion, which is associated with hyperactivity, somnolence or both. It is characterized by an acutely altered mental state, also referred to as an acute change in mental status (ACMS). We will be using delirium as the encephalopathy of interest and ACMS as a synonym for delirium.

Encephalopathy Identification

Identifying patients with encephalopathy that manifests, as delirium is important because it is prevalent, often missed\textsuperscript{2-4}, and the longer it persists the greater the risk to the patient. Importantly, some delirium is treatable. Practice guidelines for preventing and dealing with delirium have been created\textsuperscript{5,7}.

Patients who experience delirium have longer hospital stays, more ICU days, and frequently are on mechanical ventilation longer than non-delirium patients.\textsuperscript{6,8-11} They are also more likely to require post-hospital institutionalization. Studies show that patients with persistent delirium have a 2 to 4-fold higher chance of dying within the next year.\textsuperscript{8,9,12} All of this extra care is expensive. Delirium has been estimated to cost the U.S. healthcare system between $38 billion and $152 billion annually.\textsuperscript{7,13}

Delirium is estimated as being missed in the majority of older, delirious patients.\textsuperscript{2,3,14} These misses may be due to unfamiliarity with psychiatric conditions and their treatments and the prevalence of the hypoactive forms. Wancata et al. found that non-psychiatrists’ diagnostic sensitivity in detecting general psychiatric disorders ranged from 31.3-89.5\%\textsuperscript{15}. Delirium is particularly difficult because its symptoms fluctuate throughout the day. To correctly diagnose it, a clinician must look at the time course of mental status changes, the history of predisposing illness, any medications taken, any alcohol or other substance withdrawal, and pertinent environmental changes.\textsuperscript{16} It is probable that age and vision impairment are correlated with a missed diagnosis because physicians are over-looking symptoms they believe are explained by a known condition (age-related decline and perception problems, respectively).

This tendency, over-looking symptoms due to erroneous ascription, points to the need for clinical decision support. Automatically identifying documents with evidence of encephalopathy written about patients with no pertinent diagnostic codes is important to the hospital as well as the clinician and patient. Encephalopathy increases the complexity of patient care, which should be reflected in hospital reimbursement and case mix statistics.

Automated Evidence Identification
The automated identification of clinical evidence has the potential to improve clinical decision support for delirium. In cases where a physician has recognized delirium symptoms, but has wrongly ascribed them, it will be important for clinical decision support systems to identify those symptoms in order to re-direct the thinking of the clinical team. In all but a very few electronic health record systems (EHRs), the description of symptoms are found in free-text notes. ICD-9 codes and other structured text reflect current and “working” diagnoses assigned to inpatients, but symptom extraction requires natural language processing (NLP).

Our lab has developed an NLP pipeline\(^7\) to extract “disorder mentions.” Disorder mention is used in NLP to refer to all terms related to bodily malfunction. It includes signs and symptoms (sometimes grouped as clinical findings), diseases, diagnoses, and syndromes. In this paper, we will use the term “evidence” instead of disorder mention. Evidence seems a more accurate term when discussing EHR data that suggest a diagnosis because it does not imply, necessarily, that a decision has been made about what the data signify.

We have adapted the UtahPOET NLP pipeline to find evidence of encephalopathy (POETenceph) by including a realist ontology of encephalopathy.

**Ontologies**

Ontologies are used to incorporate world knowledge (e.g., semantic information) into NLP systems. An ontology includes the concepts, relationships, and rules governing their interactions.\(^8\) Most biomedical informaticists are familiar with the National Library of Medicine (NLM) Unified Medical Language System’s (UMLS) Metathesaurus and Semantic Network\(^9\), an amalgamated ontology of medical terms from multiple, diverse source vocabularies.

The UMLS Metathesaurus is not built using realist principles. Realist ontologies are designed to ensure that the concepts represented reflect real-world entities.\(^10\) The most important example of this principal from the current system is the distinction between an EHR entry and a disorder mention. If a clinician writes, “the patient has dementia and is unresponsive,” UMLS Metathesaurus mappings would be to the disorder mentions “dementia” (C0497327) and “unresponsive to stimuli” (C0857494). However, the terms “dementia” and “unresponsive” are ambiguous. Dementia could refer to Alzheimer’s disease (C002395) or any one of the other 581 related concepts in the UMLS Metathesaurus. Unresponsive could refer to “unresponsive to questioning” (C2188201) or any one of the other 35 related concepts. This term-use ambiguity is generally acknowledged. However, there is a second type of ambiguity that is conflated into the idea of a disorder mention, the clinician’s knowledge. If we assume that the sentence is linking unresponsiveness to dementia and that the clinician knows that “unresponsive to questioning” is a symptom of dementia, but “unresponsive to stimuli” is not, then we can assume the reference is to “unresponsive to questioning.” However, if our second assumption is false and the reference is to “unresponsive to stimuli,” then we should alert the clinician to their mistake in ascribing the symptom to dementia.

Realist ontology development forces us to differentiate the two types of ambiguity because we must separate what is happening in the patient, from what the clinician is thinking, and what is written in the EHR. These three things are different types of real world entities. What is happening in the patient is a level 1 entity (a real thing), what the clinician is thinking is level 2 (a mental representation), and what is written in the record is level 3 (a written representation).\(^11\)

One alternative to including an ontology is to train machine learning classifiers to classify clinical records into diagnostic categories. This method has not been generalizable across datasets\(^12\) probably because evidence is expressed in many different ways and structure of the documents differs. Therefore training a generalizable classifier would require a tremendous amount of training data. There would also be no guarantee that the resulting classifier reflects accurate biomedical knowledge. Although it would reflect the consensus of the clinicians whose documents it was trained with, there is no guarantee this is accurate.

**Ontology and Evidence Rating**

The ontology can be used to find evidence and then rate it’s association with a diagnosis of encephalopathy in the form of delirium. To find evidence is a straightforward dictionary match including lexical variants. One common way to match lexical variants is to use the UMLS Metathesaurus. We employ this technique.

To rate the association between the matched ontology concept and the diagnosis of encephalopathy, is a process of determining the semantic similarity or relatedness between the matched concept and the diagnosis. Semantic similarity is based on the separation between concepts in the hierarchical structure, while relatedness can use relationships other than hierarchical. See Batet, et al.\(^13\) and Pedersen, et al.\(^14\) for discussions of different approaches
to calculating semantic similarity and relatedness. In broad terms, there are two groups of semantic relatedness algorithms, those based on an ontology and those calculated from corpus statistics.

Corpus statistics alone may not lead to good semantic similarity judgments because the use of language cannot be used to understanding the meaning of language. It is the problem that psychologists refer to as the “symbol grounding” problem. In essence, words in a system cannot derive meaning from other words because words are arbitrary symbols with no intrinsic meaning. Using a realist ontology, we ground our words by attaching them to entities in the world. The resultant ontology is then used to judge semantic relatedness.

To make this judgment, we combine a path-length metric from the ontology with information content (IC) measures following the example of Pedersen, et al. and others. We use a simple measure of centrality for path length to determine semantic relatedness, since we have a central point representing complete knowledge (a diagnosis of encephalopathy). For IC, we used the number of term matches with and without stemming found in the UMLS Metathesaurus. IC is generally done with corpus statistics, but we reason that uniqueness within the UMLS Metathesaurus is a better indicator of uniqueness within the field. The final IC indicator we use in term length. We reason that longer terms are only included if the information they impart is necessary.

Methods

Evaluation

To evaluate our approach, we will look at two system aspects. The first is POETenceph’s ability to recognize EHR entries as evidence of encephalopathy in the form of delirium. The second system aspect will be its ability to combine the evidence to rank clinical documents based on the likelihood the patient was experiencing delirium.

We will compare the POETenceph output to 100 de-identified clinical documents from the Pittsburgh dataset. The corpus was created by author YS. The full description of the corpus creation is in Shao, et al. The documents were identified as related to delirium using topic modeling. Two subject matter experts (former nurses including co-author CW) annotated the mentions related to delirium in each document. The annotation process is also described in Shao, et al. Annotation was covered under University of Utah IRB_00043685.

We show the overall results compared to each reviewer individually and to the annotations matched between reviewers as well as the unique annotations across reviewers. With a condition as difficult to define as delirium, we think that each expert’s contribution should be considered. Therefore, for the detailed look at the POETenceph results we classified the amount of evidence based on the number of unique annotations across the two reviewers. There was a tendency for reviewer 1 to annotate only a few concepts. Documents were categorized as showing little or no evidence (up to 3 annotations or delirium denied in the text), some evidence (delirium implied in the text), evidence against delirium (another condition stated in the text), and good evidence (5 or more annotations or delirium stated in the text).

Since POETenceph is being developed to augment a word-matching system that searches for the words “enceph” and “deliri,” with proximity negation, we will also look at the occurrences of those words in the annotated corpus.

UtahPOET pipeline

The UtahPOET pipeline (Figure 1) is described in detail in Doing-Harris, et al. It is built in Apache UIMA and has the common NLP pipeline structure. Pre-processing includes sentence splitting, tokenization, and part-of-speech tagging. We add a preprocessing step for POETenceph that identifies label-value pairs (e.g., “Ca 8.6”) using regular expressions (Figure 1, section F). The regular expressions match noun-number pairs.

UtahPOET is unique in that it separates well-formed sentence (i.e., prose) from ill-formed sentences (i.e., nonprose) so that nonprose sentences can be re-split and prose sentences can be dependency parsed to facilitate processing. Non-prose sentences are split at line breaks because their internal cohesion is not important. Prose sentences are dependency parsed to allow more appropriate mapping of concepts, whose constituent words are separated in the text and to attach attributes (negation, experiencer, uncertainty, generic) to the concepts. For example, “infection in her uterus” mapped to “infection of uterus.” Generic mentions are like “psychiatry” in the sentence “patient sent to psychiatry for evaluation.”

Multi-word terms and their attributes are found in nonprose by adjacency. We reason that nonprose sentences are telegraphic and unlikely to include long-distance dependencies. Long-distance dependencies can only be understood by recognizing clausal structure.
Figure 1. The UtahPOET pipeline. Zoom in for form detail.

Post-processing (Figure 1, section K) specific to the SemEval competition is not used here. However, the removal of terms erroneously matched to disorders is retained ("sinus", "tongue", "blood", "ear", "a.m.", "dr.", "md", "dr", "D. D.", "m", "pm", "he", "mr.", "CT", "P", "Ht", "Eyes", "DATE", "T"). We also add a post-processing step in which any disorder mention found in a label-value pair is ignored. Instead of the remainder of the post-processing, we search the text of the identified disorders against the Encephalopathy ontology to determine whether they are evidence of encephalopathy.

Ontology of Encephalopathy Characterized by Delirium (OECD)

The main points of interest in the OECD are the level 3 (written representation) EHR entry, the level 2 (mental representation) clinical entities (i.e., disease picture, clinical picture, clinical finding, and diagnosis).

The upper level of a realist ontology contains the entities continuant and occurrent. Continuant subsumes the entities dependent continuant with generically dependent continuant and specifically dependent continuant and independent continuant. Explaining these entities is beyond the scope of this paper.

Figure 2. A) Proposed entities for the Ontology of encephalopathy characterized by delirium are in the red box. All other entities are from the Information Artifact Ontology. B) Proposed clinical picture entities.

Figure 2 shows some of the IAO Information content entities we propose. The EHR entry entity is a textual entity from the Information Artifact Ontology (IAO), which is an information content entity, which is a generically dependent continuant. The clinical entities we propose all fall under the Ontology of General Medical Science.
entities disease picture, clinical picture, clinical finding, and diagnosis. These entities are data items, which are also information content entities. A diagnosis of encephalopathy is a diagnosis of mental disease from the Ontology of Mental Disease. A diagnosis of mental disease is an OGMS diagnosis. The clinical findings that we propose are either OGMS clinical history findings or OGMS physical examination findings. We use the diagnostic criteria described in across several journal articles to associate clinical findings, clinical pictures, disease pictures and diagnoses. We do not attempt to address any physical aspects of the patient, only the clinician’s interpretation of the patient.

We created the ontology in stages. The first stage used automated extraction of UMLS Metathesaurus concepts from sentences hand-culled from medical records because they related to Delirium. We reconciled these concepts against the diagnostic criteria for delirium as stage two. Stage three is putting the ontology into BFO format. Stage four will include an expansion of the ontology with terms identified by our ontology management system SEAM and the terms identified in this study. The final stage will be obtaining expert agreement on the final ontology.

POETenceph

We incorporate the OECD by creating two Apache Lucene indices. The first index contains the unique ontology identifiers and labels (i.e., ontology key) associated with each of the ontology entities. Centrality scores are calculated for each entity by starting with the diagnosis (encephalopathy) and associated disease pictures (encephalopathy and urinary tract infection), which are set to centrality of -1. As you travel out across relationships from the disease picture, each level adds 1 to the centrality. The scores are normalized with division by the longest path length. Example centrality, diffusion, and length scores are listed in Table 1.

Table 1. Example centrality, diffusion, and length scores for OECD entities.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Centrality</th>
<th>Diffusion</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delirium</td>
<td>0.0</td>
<td>1 - 0.002</td>
<td>0.1</td>
</tr>
<tr>
<td>Encephalopathy</td>
<td>0.0</td>
<td>1 - 0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>Mental Disease</td>
<td>0.875</td>
<td>1 - 0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>UTI</td>
<td>-0.125</td>
<td>1 - 4.2 E-4</td>
<td>0.3</td>
</tr>
<tr>
<td>Memory Impairment</td>
<td>0.625</td>
<td>1 - 5.6</td>
<td>0.2</td>
</tr>
<tr>
<td>Hyperactive delirium</td>
<td>0.625</td>
<td>1 - 6.3 E-6</td>
<td>0.2</td>
</tr>
</tbody>
</table>

We designed the second index to accommodate the variability in language used by clinicians and to measure diffusion. It was created by matching all of the terms in the ontology (EHR entry instances and entity labels) to the UMLS Metathesaurus. Matches were restricted to the clinical source vocabularies and disorder or medication semantic types. In this index each term was associated with the ontology key of the search term that returned it, and its measures of diffusion, centrality, and length (see Table 1 for examples). Diffusion was the number of matched terms returned. Once all the terms have been searched the maximum number of matches was used to normalize the diffusion of each term, keeping it between 0 and 1. The result is subtracted from 1 because less diffuse is better.

The final aspect of the term that we used to calculate its score was the length of the term again normalized by the length of the longest term. The final score is the mean of the centrality, diffusion, and length. We score the document with the sum of the scores for the evidence that is positively asserted (not negated), experienced by the patient, and not generic.

Results

The most important test of POETenceph is determining whether high rated and low rated documents are correctly categorized. In the first set of figures we show two ways of looking at the human annotations across the documents. Figure 3a shows the POETenceph scores as they compare to each reviewer individually. Figure 3b shows the POETenceph scores compared to the annotations matched across the reviewers and the unique annotations found when combining the responses of the two reviewers. In both images it can be seen that the trend of the POETenceph scores follow those of the reviewers. It is apparent that they follow reviewer 2 more closely than reviewer 1 and the unique annotations better than the matched annotations. We find this preference for reviewer 2, who also led the number of unique annotations, was due to her annotating a higher number of pieces of evidence. Visual inspection
indicated that her annotations were valid. In one instance (file #44), she annotated “Change in mental status,” “sluggish,” and “combative” that reviewer 1 did not.

![Figure 3](image1)

**Figure 3.** Comparison of scores from POETenceph and Reviewer results: A) Each reviewer separately. B) The annotations matched between the reviewers and the unique annotations across the two reviewers.

In a more detailed examination, we looked at the top 20 and bottom 20 documents for the amount of evidence highlighted by the human reviewers (see Figure 4). In the bottom 20, 13 showed little or no evidence of Encephalopathy or Delirium (65%) within this group 1 had controlled encephalopathy and 1 had controlled delirium. One document showed some evidence of delirium (5%), and 6 showed good evidence (30%). The 20 highest rated documents showed almost the same pattern 14 had good evidence for delirium (70%), 3 had evidence of another condition (15%), 1 had controlled delirium (5%) and 2 had little or no evidence (10%). Three high scoring documents had evidence for related conditions (depression, alcohol withdrawal, and kidney disease), but without delirium.

![Figure 3](image2)

**Figure 3.** Comparison of scores from POETenceph and Unique Annotations: A) the bottom 20 files. B) The top 20 files. ○ indicates the text states no delirium, ★ indicates that the text states delirium present.

Looking at the word search for “enceph” we find only 11 documents contained the string, 3 from the low scoring documents 3 showing little evidence and 1 showing some evidence. Four of the high scoring documents contained the string “enceph,” 3 had good evidence of delirium and 1 had some evidence.

Matching the string “deliri” we find more coverage. Sixty-four of the 100 documents contain the string. It is found in only 9 of the 20 lowest scoring documents and only 4 of those contain little or no evidence, 3 contain good evidence and 2 contain some evidence. Of the 20 highest scoring documents 17 contain the string “deliri.” Two of the mentions contain little or negated evidence, 2 contain other conditions with delirium stated as not being present. Two high scoring documents with good evidence for delirium did not contain the string.
We then compared the POETenceph annotations with the expert annotations. We again looked at 21 documents across a range of scores from low to high. We found two obvious problems because we built UtahPOET to identify disorder mentions, POETenceph does not find medication or instrument mentions. These entities exist in the ontology but comparison to the ontology is done after UtahPOET has run.

Looking only at the direct evidence for delirium, POETenceph consistently misses “mental status changes,” “unable to follow commands,” “unresponsive,” “agitated,” and “not communicate.” Mental status changes, unresponsive, and agitated are simply missed by UtahPOET. Unable, follow and communicate are identified as concepts by UtahPOET, but not as evidence by POETenceph. The other class of evidence that is missed is evidence that is not found in the UMLS Metathesaurus. For example, “screams alternating with drowsiness,” “minimally verbal,” or “not clear mentally” are not found there.

POETenceph identified 249 positively asserted pieces of evidence, 64 are agreed-upon by at least one reviewer (26%). Eighty-six (35%) annotations are problems from UtahPOET, including repeated erroneous mappings of “M],” “man,” “block,” “he” (12 times), and “ca” (16 times). The remaining 98 we think are related to encephalopathy, but missed by the reviewers either because their task was to locate direct evidence (39%). This task orientation may also explain why only 5 of the 30 useful negated pieces of evidence were agreed-upon by a reviewer. POETenceph erroneously negated five others. Twenty-five of the negated evidence findings were errors like those found in the positively asserted evidence and therefore not useful.

The annotators found 107 pieces of evidence missed by POETenceph. Twenty-one were medications or instrument mentions, 3 are not disorders (Psychiatry, ammonia, potassium), 47 were pure misses (e.g., “agitated,” “acute psychotic event,” and concepts including “mental status”) and 38 are not in the UMLS Metathesaurus.

**Discussion**

We found that the POETenceph performed the task of ranking documents based on the amount of evidence of encephalopathy in the form of delirium well. Higher scores indicated documents that contained more reviewer annotations. The scores tracked particularly well with reviewer 2 and the total number of unique annotations in a document. Seventy percent of the top 20 and bottom 20 scoring documents were appropriately placed. Only 15% of the documents in top group and 20% in the bottom were completely misplaced. These findings indicates that using a realist ontology combined with measures of centrality, diffusion and length, it is possible to calculate scores for evidence in a document that corresponds well with expert judgments.

By restricting our search to documents without diagnostic codes for encephalopathy, we will be able to identify evidence that should be brought to the attention of clinicians and medical coders. This evidence can be presented to clinicians to encourage them to re-evaluate the patient or the billing codes whichever is appropriate.

This performance is an improvement over string search. The string “enceph” only returns 11 of the documents two of which would be completely misplaced. The string “deliri” returns many more of the documents because it was used in the creation of this document corpus, but the documents are unranked so although only 6 contain little or no evidence there is no indication of which 6. Also 3 documents with good evidence would be missed.

This study alone does not prove the utility of realist ontologies. It does indicate that they are useful. Having an ontology for the construction of a diagnosis from clinical findings that is separate from the UMLS Metathesaurus allows us to exploit the Metathesaurus’ ability to represent term use. We can also change the diagnostic elements without advocating for changes to the Metathesaurus.

The most obvious problem exposed in this study is that UtahPOET is restricted to matching disorder mentions whereas clinical evidence also comes from medication and instrument mentions. There are other improvements to UtahPOET that would aid performance of POETenceph. We need to address the erroneous mapping of small strings like “he” and “M],” as well as the missed mappings of “agitated,” “mental status” concepts, and “depressive.” Our next step will be to address the problems with UtahPOET.

In another project we are working on formalizing the Encephalopathy ontology. This formalization will include fixes to the current associations between the entities as well as the extension into level 1 entities. The formalization process requires input from clinicians to determine the settled science linking level 1 entities, the state of the practice linking level 2 entities and the imprecision of clinician language when transcribing their thoughts and impressions into the EHR. We will be investigating the use of probabilities on the relation between EHR entries and mental representations. Once these extensions have been added to the ontology we will need to reexamine how we calculate centrality.
For concepts not found in the UMLS Metathesaurus from the SemEval challenge we developed a structure SVM component for UtahPOET. We did not have training data to train the component for the equivalent pieces of evidence nor did we try the component with its current training because it does not currently work well. In future work, we will upgrade the component and obtain more training data for it.

The final area of investigation that this study has spurred is the relationship between disorders and body locations. Currently we determine each of these concepts separately. However, there must be a method to use one as a sanity check against the other. It seems that if a body location is found it should be used to restrict the possible matches to its attached disorder. For example, and that some disorders carry a body location with them “angina pectoralis” is only in the chest. We may be able to extend this idea to the mapping of strings into the ontology.

In future work we will also investigate the balance of the three measures, which go into our scores. It is possible that the current balance gives terms length too much impact on the final score for its informational content. It may be better to weight centrality higher than length and possibly also diffusion.

Limitations

The test corpus had only documents that had been previously identified as relevant to delirium. Therefore, we may have our encephalopathy detection set too loose. Because we are in no danger of missing delirium mentions, we cannot be sure we would catch them all. Our next evaluation will contain a predetermine mixture of cases.

We are also limited in that we did not enforce agreement between the annotators. We do not believe that it is necessary for all expert reviewers to agree on each and every annotation. Individuals will always have a slightly different opinion on what constitutes important evidence for conditions as difficult to define as delirium. Forcing arbitrary agreement would create the illusion of a strict definition. We do think that having several reviewers would allow us to choose annotations favored by the majority. A majority opinion may be more indicative of the prevailing medical consensus. In future work, we will strive to have five reviewers for each file, although that will have implications for the number of files we will be able to have reviewed.

Conclusion

An effective document ranking system can be created by combining a disorder-identification NLP pipeline with a realist ontology. Comparing disorders to ontology entities and scoring the matches based on diffusion, centrality and the term length can assess the evidentiary value of disorder mentions.

Acknowledgements

This work was supported in part by a grant from the NLM, R01-LM010981. This material is based upon work supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Biomedical Laboratory Research and Development: Veterans Health Administration Health Services Research & Development: # CRE 12-321.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

References


Strategies for Managing Mobile Devices for Use by Hospitalized Inpatients

Patricia C. Dykes PhD, RN1,2, Diana Stade1, Anuj Dalal MD1,2, Marsha Clements, PhD(c), RN1, Sarah Collins, PhD, RN2,3, Frank Chang MSE3, Anne Fladger, MLS, AHIP1, George Getty3, John Hanna1, Raval Kandala3, Lisa S. Lehmann MD, PhD1,2, Kathleen Leone MBA, RN1, Anthony F. Massaro MD1,2, Eli Mlaver1, Kelly McNally1, Sucheta Ravindran1, Kumiako Schnock PhD, RN1, David W. Bates MD, MSc1,2,3

1Center for Patient Safety, Research and Practice, Brigham and Women’s Hospital, Boston, MA, 2Harvard Medical School, Boston, MA, 3Partners HealthCare, Boston, MA,

Abstract

Despite the potential advantages, implementation of mobile devices and ongoing management pose challenges in the hospital environment. Our team implemented the PROSPECT (Promoting Respect and Ongoing Safety through Patient-centeredness, Engagement, Communication and Technology) project at Brigham and Women’s Hospital. The goal of PROSPECT is to transform the hospital environment by providing a suite of e-tools to facilitate teamwork among nurses, physicians, patients and to engage patients and care partners in their plan of care. In this paper, we describe the device-related decisions and challenges faced including device and accessory selection, integration, information and device security, infection control, user access, and ongoing operation and maintenance. We relate the strategies that we used for managing mobile devices and lessons learned based on our experiences.

Keywords: device management, communication, health information technology, patient engagement, nursing informatics.

Introduction

The use of technology to manage health and wellness is becoming commonplace. Adoption and use of web-based technologies in everyday life has exploded over the past decade1,2 and the majority (87%) of American adults use the internet1. Many internet users go online to get health information for themselves (72%), or for a family member or friend (50%)1. The use of mobile devices is also becoming widespread; 90% of Americans own a cell phone, over half (53%) a smart phone,3 and increasingly, many own tablet computers. Many Americans report that they use their mobile device as a tool for managing their health. A recent article indicates that there are approximately 90,000 mobile health apps to manage a range of health-related activities1.

Many consumers are using technology to manage health and wellness, and there is a growing recognition of the need for patient engagement in healthcare5. The Health Information Technology for Economic and Clinical Health (HITECH) Act meaningful use program (2009) requires that providers engage patients in their healthcare through the use of technology. A key goal of meaningful use is to make consumers full partners in their care by providing e-health tools that increase access to health information, support activation (e.g., active involvement in their treatment plan), and that help consumers to gain control over their health and wellbeing6,7. This has led to provider and healthcare organizations promoting the use of personal health records or patient portals in outpatient settings8,9. However, there are limited examples in the literature that describe strategies and e-health tools to provide patients with access to their health information in hospital or inpatient settings. Based on a systematic review of patient engagement
technologies in inpatient settings, Prey et. al. (2014) reported a dearth of published papers describing technologies to engage patients in inpatient settings. The research that has been done in this area is largely confined to pilot and usability studies. None of the papers included in the systematic review described in detail the operational issues related to implementing e-health tools and devices to engage patients, the infrastructure needed for everyday use, and the management issues associated with implementation of these tools in busy acute care settings by a multitude of patients. Early work by Berg et. al. (1999) argue that successful design and implementation of health information technology (IT) applications requires a socio-technical approach; e.g., insight into specific work practices and the dynamic interaction between users and technology in the context of their workflows

In 2014 our team from the Center for Patient Safety, Research, and Practice at Brigham and Women’s Hospital (BWH) implemented the PROSPECT (Promoting Respect and Ongoing Safety through Patient-centeredness, Engagement, Communication and Technology) project on BWH medical intensive care (MICU) and oncology units (http://www.partners.org/cird/PROSPECT/Index.htm). PROSPECT is funded by the Gordon and Betty Moore Foundation and BWH. The units (20 ICU beds, 40 oncology beds) were selected to participate based on the GBMF’s interest in targeting ICUs and the relatively large percentage of oncology patients that are transferred to the MICU. Our goal is to transform the acute care environment by providing a suite of e-tools to facilitate teamwork among nurses, physicians, patients and to engage patients and their care partners in their plan of care. We hypothesize that through this partnership, we can eliminate preventable harms, make health care more affordable, ensure that patients and families make informed choices about their health care, and improve communication and concordance on goals of care. One component of the PROSPECT project is a Patient-centered Toolkit (PCTK). The PCTK is a web portal that is designed to provide patients and their care partners with the core set of information needed to engage in their plan of care during an acute hospitalization. In addition, patients can use the PCTK to message their care team and to provide feedback on their plan of care.

As part of the PROSPECT project, we partnered with our medical librarian to research strategies for ongoing operation and maintenance of bedside devices used by patients in hospital settings and found no peer reviewed literature on this topic. After implementing mobile tablets at each bedside for BWH MICU and oncology patients, we have learned some important lessons. In this paper, we focus on strategies for implementing and managing mobile devices in hospitals and lessons learned based on our experiences from the PROSPECT project. Key questions related to bedside device implementation and management are included in Table 1.

<table>
<thead>
<tr>
<th>Device Selection</th>
<th>Device Management</th>
<th>Device Related Policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>• What are the technical requirements for bedside devices?</td>
<td>• How will devices be stored (e.g., at the bedside or in central location)?</td>
<td>• How will patients be enrolled?</td>
</tr>
<tr>
<td>• What are the end-user hardware requirements for bedside devices?</td>
<td>• How will devices be cleaned?</td>
<td>• How will care partners be enrolled?</td>
</tr>
<tr>
<td>• What accessories are needed?</td>
<td></td>
<td>• What is the process to ensure compliance with institutional policy, IRB (informed consent) and HIPAA regulations?</td>
</tr>
<tr>
<td>      − Will patients need a tangible keyboard?</td>
<td></td>
<td>      − How will user names and passwords be issued?</td>
</tr>
<tr>
<td>      − Are device covers/cases needed?</td>
<td></td>
<td>• How will information security requirements be addressed?</td>
</tr>
</tbody>
</table>

Table 1: Key Questions Related to Bedside Device Implementation and Management for the PROSPECT Project
Methods

We used a socio-technical systems approach\textsuperscript{14} to develop the PCTK\textsuperscript{12}, to identify device requirements, and to identify strategies for ongoing device management. In the 10 months prior to PCTK implementation, we used a range of strategies to gain insight into the MICU and oncology units’ work practices into which we would implement the PCTK. We conducted workflow observations, interviews, and focus groups of care team members, patients, and care partners. Based on our evolving set of requirements and the concerns expressed by clinicians and patients, we scheduled meetings with hospital, health information management (HIM), and infection control leadership to address the key questions that arose (Figure 1). In addition, we met regularly with information systems (IS) leadership to negotiate integration with existing BWH and Partners HealthCare (PHS) IT systems and to secure approval for device integration and management plans. The details of the iterative, user-centered PCTK development approach are described elsewhere\textsuperscript{10-13}. Here we focus on the on the socio-technical aspects of identifying device requirements, the procedures needed for managing bedside mobile devices for the PROSPECT project, and our experiences to date.

Results

Using the socio-technical systems approach, we identified challenges in terms of device and accessory selection, user access, integration, information and device security, infection control, and ongoing operation and maintenance in the hospital environment. In the sections below we describe the requirements and the implications related to implementing mobile devices at the bedside for patient use.

1. Device type, accessories, storage and user access

Based on initial feedback from patients and clinicians, we had originally planned to install touch screen, interactive, “smart” devices with built-in keyboards on a moveable arm at the bedside. These devices would support patient access to the PCTK and could be stored flat against the wall when not in use. However, we were unable to implement this approach due to space constraints in the room and the organizational decision to implement a non-touch screen device for use by clinicians in each patient room (device would not be accessible to patients). Therefore we made a decision to use mobile tablet devices. One implication of this decision was that our programmers were developing the PCTK software for use on a Linux, Windows\textregistered operating system and a selection of a device that ran on a different operating system, such as Android or iOS, had development implications.

Once the decision was made to use mobile devices, several issues required consideration including mobile device selection, whether to provide a keyboard, and identification of a device storage strategy that would promote patient and care partner access throughout the hospitalization and ensure the device could be easily charged. We initially identified the following bedside device requirements based on the PCTK technical requirements and user feedback: light weight, optimized screen size, multi-touch screen capability, Bluetooth low energy, and virtual keypad functionality. We explored multiple mobile devices and notebooks and rated each including the iPad Air, iPad mini, Nexus 7 and 10, Surface Pro, ATIC PC Smart, Xperia Series, the Galaxy Note 10.1, Arrows Tab, Macbook Air, LaVie Series, and the CF-LX3. After consideration of infection control (notebook keyboards were difficult to clean) and storage requirements (see below) we made the decision to go with a mobile tablet, rather than a notebook. The Apple iPad Air was the best fit as it met technical and end user requirements, required minimal instruction for use, and received positive feedback for usability from end users. In addition, PHS was using iPads
with patients in other projects and our development team had experience programming on the iOS platform. We purchased 50 iPad Air 2, Wi-fi, 16 GB devices (Apple, Cupertino, CA) for use by patients (and/or designated caregivers) on PROSPECT units to access the PCTK. The total cost for the devices (including the warrantee) was $37,900 ($758 x 50 devices).

We also decided to enclose all devices in a case. Key requirements for device cases were ease of decontamination between patients (no crevices) and preventing damage if the device was dropped. The case needed to be easily gripped, durable, waterproof, and lightweight. We reviewed two different cases including the Trent Airbender 1.0 and the military grade Griffin Survivor Case. The Trent Airbender included a keyboard but it was unclear how the keyboard could be cleaned between patients. We selected the Griffin survivor case which was commercial grade and met all of our requirements. The total cost for the cases was $3999.50 ($79.99 x 50 cases). A small percentage of patients we interviewed stated that they needed a tangible keyboard for use with the iPad so we decided to provide accessory keyboards upon request. The keyboards we chose were flexible bluetooth/wireless waterproof keyboards. To date, we have not received any requests for the accessory keyboard. The total cost for the keyboards was $23.98 ($11.99 x 2 keyboards).

The main requirement for device storage and user access was that patients have easy access to the PCTK at all times. This meant that each room would have its own device and charger that are treated as part of the room’s equipment and be within reach of the patient from hospital bed. While iPad stands and wall brackets were considered, they were not selected. iPad stands raised concerns about patient falls and wall brackets were ruled out because they did not enable the patients and care partners to use the iPad from anywhere in the room (e.g., the bed, a chair, or couch). Based on bedside workflow observations on inpatient Oncology and MICU units different device housing was developed for each environment. The rooms in the MICU were smaller and had more equipment than the rooms on the Oncology units. Both environments had over-bed tables but the tables did not have the capacity or potential to hold a mobile device in a practical and sustainable way. We looked at other existing equipment in the rooms and found that in Oncology there were bedside nightstands. Plastic bins (see Figure 1a) were purchased and installed on each side of the bedside table to facilitate access from the bed or the chair.
unit director, baskets were emptied and the devices and chargers were placed in the baskets upon initial installation (see Figure 1b). Over time, the baskets were filled back up with supplies in addition to the devices. Meanwhile, our second and long-term solution, recommended by the clinical staff, was to use IV poles with baskets on them which could be placed near the bed but easily wheeled to the side if the space was needed for bedside care (see Figure 1c). We installed the poles in each of the rooms and placed the devices in the baskets. Over a few weeks, the poles were either pushed out of the rooms by clinicians, or, they were used for intravenous therapy and not for housing the devices. The clinicians stated that they did not like the extra equipment in the room. Therefore, they have reverted back to using the baskets on the wall to hold the devices. While not ideal, this solution was acceptable because the patients in the MICU are often times incapacitated and it is more common that the Health Care Proxy/care partners use the PCTK. Therefore access from the bed is not always needed.

2. Device integration strategy and device security

We worked with our development team and PHS IS department to identify a device integration strategy that met the needs of the PROSPECT project and was consistent with PHS security policies. First, patients and care partners needed to access personal health information and communicate about their plan of care with the care team; therefore, all devices were connected to the PHS secure wireless network. Second, PHS policy required passwords to be centrally managed and information on devices to be cleared after each patient use. We used an enterprise mobile device management (MDM) solution (AirWatch® by VMware®, Atlanta, GA) to monitor the 50 mobile devices used by approximately 1000 patients/care partners. The MDM software provided remote security controls and reporting capabilities that met both PHS information security policies and HIPAA regulations. Specifically, the MDM software ensured that the iPads were assigned to users, compliance policies were managed, and device tracking was done from a central console. Thus, any device could be quickly locked down and secured when necessary.

<table>
<thead>
<tr>
<th>Device and WiFi Issues</th>
<th>Number of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chargers missing/replaced</td>
<td>30</td>
</tr>
<tr>
<td>iPads stopped working properly (touch screen functionality failed)</td>
<td>2</td>
</tr>
<tr>
<td>iPad disabled from too many logins</td>
<td>4</td>
</tr>
<tr>
<td>iPad missing</td>
<td>1</td>
</tr>
<tr>
<td>Documented WiFi issues</td>
<td>3</td>
</tr>
</tbody>
</table>

*Table 2: PROSPECT Project device and WiFi issues July 2014-February 2015 (N= 50 iPads; 305 patients/care partners)*

We also used content management software, Kiosk Pro Lite (Kiosk Group, Inc., Frederick, Maryland) to restrict access on the iPad to the PCTK only. The iPads were configured by the MDM application to only allow access to the Kiosk Pro Lite software. This prevented data from being downloaded on mobile devices, in compliance with PHS information security policies. All personal health information and patient-provider messaging functionality on
the PCTK were available via secure web services. Therefore, we eliminated the need to clear data on devices between each patient use (although with the AirWatch software, we did have this capability).

Prior to implementation, we were concerned about device theft. We mitigated this threat not only by restricting user access to the PCTK, but also by configuring the MDM software to disallow internet access on all devices outside the PHS firewall. Therefore, the iPads were rendered useless if stolen. We communicated this to patients and clinical staff via signs and weatherproof/laminated, high-tack, white, vinyl labels on the back of the iPad case (see Figure 2). These labels were also applied to each of the chargers and cords. Thus far, our approach has been successful. Since go-live in July of 2014, only one device has been reported as missing.

![Figure 2: iPad Security Label](image)

We identified issues related to devices and WiFi connectivity after implementation of the PROSPECT project (see Table 2). Locking down the iPad devices in kiosk mode did have unintended consequences in that we were unable to leverage the full interactive capabilities of the iPad. First, all pop-ups were suppressed; we were unable to use iframes for some external links or third party apps. This limited the content that we could display. Second, we were unable to use the iPad’s notification functionality to display alerts on the lock screen (e.g., when receiving a message from the provider). Finally, patients and care partners could not use the iPads for personal web-browsing, entertainment, email, or video.

Because PROSPECT is a research project, we needed an enrollment strategy that met IRB requirements, PHS information security policies and HIPAA regulations. The IRB required informed consent from all patients and care partners. Patients who had capacity could provide permission for one or more care partners to access their personal health information via the PCTK. For patients without capacity, the healthcare proxy could be enrolled and could use the PCTK on the patient’s behalf. First, the process involved securing informed written consent from the patient or health care proxy using the 8-page consent form approved by the IRB. We adopted portal access best practices where possible including providing unique user names and passwords for each user and following a defined permission process. We leveraged existing BWH/PHS policies and practices related to enrollment in the PHS enterprise outpatient portal, Patient Gateway. Once consented, patients/health care proxy completed a PHS access authorization form. An additional form was completed by the patient and care partner for each care partner that the patient identified. Next, we used the PHS administrative portal to enroll patients and care partners. Specifically, users were assigned unique user names (an existing email address) and passwords (assigned by administrative portal and changed by user at the time of the first login).

3. Infection control procedures

During interviews before implementing the PROSPECT project, both patients and clinicians expressed concerns regarding device cleaning procedures. We met with hospital infection control specialists to establish a systematic process for disinfecting devices between patients. To decrease the risk of contamination associated with devices...
being carried in and out of rooms, we assigned one device to each patient room. We communicated the policy that all devices stay in the rooms at all times to hospital staff, patients and care partners. We also posted signs in the patient rooms reminding patients and care partners that the devices were for use in the patient room only. We worked with the BWH environmental services team to implement the approved process for cleaning both iPads and device holders after each patient use along with the other standard equipment in the patient room (see Figure 3).

Figure 3: Process for Disinfecting iPad Holders Between Patients

Step 1: Don gloves.
Step 2: Remove iPad from plastic holder.
Step 3: Wipe all of the exposed surfaces of the plastic device holder (outside and inside) using standard disinfectant wipes in use on BWH patient care units. The surface must be wet for the designated time listed for the disinfectant wipe.
Step 4: Check for any leftover stains and repeat as necessary.

Process for Disinfecting iPads between Patients

Step 1: Don gloves.
Step 2: Unplug all connectors from the device and remove any visible stains using standard disinfectant wipes in use on BWH patient care units.
Step 3: Wipe front of iPad using the standard disinfectant wipes in use on BWH patient care units starting at the top corner moving left to right from the top of the device to the bottom of the device. The process of cleaning the front of the device should take no less than 10 seconds. The surface must be wet for the designated time listed for the disinfectant wipe.
Step 4: Wipe top of frame and proceed by rotating the iPad in 90 degree steps until all sides of the frame have been cleaned.
Step 5: Turn the device and wipe the back surface starting at the top corner moving left to right from the top of the device to the bottom of the device. The process of cleaning the back of the device should take no less than 10 seconds. The surface must be wet for the designated time listed for the disinfectant wipe.
Step 6: Check for any leftover stains and repeat as necessary.


Discussion

Despite the advantages, the process of implementing and managing mobile devices in the hospital setting poses multiple challenges. Using a socio-technical approach, we identified a host of issues related to accessory selection, user access, integration, information and device security, infection control, ongoing operation and maintenance. Overall, the socio-technical approach has been useful for identifying and addressing device-related issues and concerns with stakeholders as part of the project planning process. The mobile devices (iPads) have been available on our MICU and Oncology units for over eight months; thus far, the experience has been positive. Enrolled patients and care partners have used the PCTK via hospital-issued iPads to communicate with the care team, contribute to their plan of care, and access their personal health information. To date, we have experienced minimal technical issues with the devices. While we were concerned about device theft prior to implementation, theft of the iPads has not been a problem. Interestingly, missing chargers has been the most common problem, one that we have not solved. Despite our efforts to include stickers on the chargers that indicate that they are BWH property, we have replaced 30 chargers so far (at $19.00 for adaptor and $29.99 for 6ft cord equates to $1,469.70 or an additional $200
per month in charger replacement costs). We are currently working on a secured solution for managing the device chargers.

We learned throughout the project that device-related decisions can have unintended consequences. As noted, we made the decision to use mobile devices after we learned that we could not install interactive, “smart” devices on a moveable arm at the patients’ bedside. This decision required some rework by our software developers to optimize the PCTK for use on the iOS operating system. Additionally, we had to address mobile device storage, charging, infection control, and security issues within our original timeline. Using a socio-technical approach, we worked through these issues with stakeholders and arrived at solutions in parallel with the PCTK software development. Some of the solutions were more successful than others. For example, establishing a device storage protocol so that devices were accessible by patients, continuously charged, and not in the “way” of clinical staff was difficult. We found that there was not a single best way to meet these requirements across units and we implemented two different solutions, one more successful (oncology units) than the other (MICU).

Our decision to lock iPads into kiosk mode facilitated information and device security, but it also led to an unexpected loss of access to MedlinePlus content. Initially, we linked to MedlinePlus educational content for test results, medications, and problems. We chose MedlinePlus because the content was at a consumer level of health literacy, had broad content coverage, and did not have advertisements. MedlinePlus contained licensed content from other sources and their license agreements did not permit framing of their content from their site. Medline Plus started blocking the framing of educational content in December 2014, several months into our project and our patients lost access to all of the educational content (e.g., all of the infobuttons “broke”). We quickly changed to other content sources and we started the process of implementing the web service version of MedlinePlus Connect so that we can once again use that content.

Our information security requirements have been a challenge at times. PHS requires encryption of all mobile devices that display patient health information including a lock screen and passcode. The lock screen appears when the battery power is less than 20% or if the device is restarted (e.g., the battery completely dies and the device is then plugged in). Even though the devices are in kiosk mode, a separate passcode is required to bypass the lock screen. When the lock screen appears, it can only be unlocked with the assistance of our research staff.

Cost represents a barrier to widespread adoption of mobile devices. In addition to the hardware and software costs, the administrative infrastructure needed to keep track of the devices and to enroll patients is significant. For the PROSPECT project, devices were managed by the research team. Sustainability outside the context of a research project will require an organizational approach to systematically address the issues identified by our team. Building this infrastructure would require institutional commitment and resources. One lower cost strategy that would mitigate access, device security, infection control, and operational concerns is to allow patients and care partners to access the PCTK on their own mobile device. A “bring your own device (BYOD)” strategy has much appeal as hospitals are essentially liberated from concerns related to removing prior data and configuring access to device after each patient use, monitoring and tracking devices to prevent theft, cleaning devices per hospital infection control protocols, and maintaining devices and accessories (e.g., losing chargers). A BYOD strategy should improve accessibility from the locked screen as patients and care partners would not have to remember additional passwords. To ensure privacy and security of patient health information, a BYOD strategy would require access via web-based patient portal applications with standard username and password authentication requirements. The use of native apps downloaded onto the mobile device to access web-based patient portal applications should enhance interactivity (i.e., by leveraging device specific API’s such as push notifications), as long as patient data resides on secure, HIPAA-compliant servers. Ideally, patients and care partners would download an app onto their mobile device at their convenience (e.g., as part of the hospital admission process). Additionally, such a strategy may increase flexibility for access by care partners who are unable to be at a patient’s bedside. Finally, with a BYOD strategy providers could “prescribe” mHealth apps that serve to engage patients in understanding and managing their care during recovery and after leaving the hospital16. However, this requires a rigorous mHealth app certification and
approval process to be in place. Next steps for the PROSPECT project include developing a BYOD strategy that will improve patient access to the PCTK on all BWH inpatients units.

There are several limitations associated with this project. We implemented the devices at a single hospital and within the context of a research study. It is not known how generalizable our experience is to other hospitals or healthcare settings. In addition, we have been using research assistants to enroll patients and track the devices. More research is needed to evaluate how these device-related functions could be integrated into a hospital environment without this level of support.

To our knowledge, this is the first report of lessons learned related to implementing mobile devices at the bedside for an extended period of time. We hope that with this paper, we have started the conversation and that others will publish their experiences so that barriers to implementation of e-health tools will be identified and addressed and these tools can be used more widely to engage hospitalized patients.

Acknowledgements

The authors would like to thank the BWH patients, family members, nurses, and physicians who provide ongoing feedback to inform our mobile device strategy. The BWH PROSPECT Project is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation.
References

Introduction

The US healthcare system is highly fragmented and fraught with attendant inefficiencies in delivery. Lack of interoperability and care coordination have resulted in duplication of care, increased error rates, adverse drug-drug interactions, reduced safety, and increased costs. It has been argued that investments in health information technologies will radically transform the healthcare sector by increasing efficiencies, decreasing expenditures and increasing quality. Prevalence of chronic diseases, and the need for improved quality of care and patient outcomes necessitates the application of Health Information Technology (HIT) and Health Information Exchange (HIE) to streamline patient care, eliminate waste, and improve care coordination, with the goal of ultimately improving patient health outcomes. This study aims at exploring the effects of HIT/HIE on the outcomes and quality of care for diabetes, asthma, acute myocardial infarction, congestive heart failure, and hip and knee fracture. We look at average length of stay, adverse drug events and 30 day readmissions as measures of quality of care. This study addresses the policy implications and areas for further research to increase the success of HIT implementation, success of HIEs and ultimately improve health care outcomes and reduce health care expenditures through improved care coordination across disparate provider entities, and readily accessible patient data across fragmented networks.

Methods

This study uses the Health Information Management Systems Society (HIMSS) IT database, coupled with the American Hospital Association (AHA) Hospital survey data and the Centers for Medicare and Medicaid (CMS), outpatient and skilled nursing facility Medicare claims data from 2006-2012, along with the standard Medicare file of patient claims for 2006-2012 as the control cohort. Using hierarchical clustering analysis methods, we construct HIT scores for measures of HIT and HIE implementation and penetration across 32 measures of HIT adoption. CMS, AHA and HIMSS data, is combined using the common Medicare Provider Number. The scores are then used as the coefficient of interest using fixed effects methods and multinomial regressions, controlling for hospital and patient characteristics, as well as time-invariant factors.

Results

This study finds that those institutions with the highest HIT scores for specific measures of clinical decision support, continuity of care documentation and clinical discharge and summary care documentation showed modest and statistically significant levels of improvement in health care quality outcomes for the 3 key outcome metrics of over the study period. Overall, more intensive use of HIT/HIE in the long run could help providers achieve better quality outcomes.

Discussion

This study has major implications for the future of HIT/HIE in coordinating care across disparate entities. Federal government initiatives aimed at improving care coordination, fostering health ecosystems, and reducing costs associated with unnecessary readmissions and length of stay and adverse drug events stand to benefit from this research in that it highlights tangible areas where HIT/HIE participation may be beneficial to achieving improved health outcomes and reduction of cost.
A Flexible Simulation Architecture for Pandemic Influenza Simulation

Henrik Eriksson, PhD, Toomas Timpka, MD, PhD, Joakim Ekberg, PhD, Armin Spereo, Örjan Dahlström, PhD, Magnus Strömgren, PhD, Einar Holm, PhD
1 Dept. of Comp. and Inform. Sci., Linköping University, Sweden;
2 Dept. of Medical and Health Sci., Linköping University, Sweden;
3 Dept. of Social and Economic Geography, Umeå University, Umeå, Sweden

Abstract

Simulation is an important resource for studying the dynamics of pandemic influenza and predicting the potential impact of interventions. However, there are several challenges for the design of such simulator architectures. Specifically, it is difficult to develop simulators that combine flexibility with run-time performance. This tradeoff is problematic in the pandemic-response setting because it makes it challenging to extend and adapt simulators for ongoing situations where rapid results are indispensable.

Simulation architectures based on aspect-oriented programming can model specific concerns of the simulator and allow developers to rapidly extend the simulator in new ways without sacrificing run-time performance. It is possible to use such aspects in conjunction with separate simulation models, which define community, disease, and intervention properties. The implication of this research for pandemic response is that aspects can add a novel layer of flexibility to simulation environments, which enables modelers to extend the simulator run-time component to new requirements that go beyond the original modeling framework.

Introduction

Computer-based simulation is an important resource for analyzing the dissemination of novel infectious disease agents before empirical data are available [1]. Such simulations can provide valuable information about the relative impact of different potential interventions measures and estimate the relative susceptibility in different types of communities [2]. A major challenge for simulation architectures to be used in these settings is to provide sufficient turn-around time for simulation runs. In addition to the run-time performance per se, the simulation architectures must provide sufficient flexibility to allow for changes to both the simulation model and to the simulation performance engine. As in many other types of software systems, there is a tradeoff between run-time performance and modeling flexibility. Hard-coded and hand-tuned simulator implementations can take advantage of available hardware, such as multi-core processors, and can achieve acceptable or even respectable performance. It is a significant challenge to combine performance and flexibility to attain a well-balanced simulation system that can support interactive modeling, scenario development, and simulation result analysis for research and decision-support tasks.

Previously, we have studied stimulation architectures that take advantage of ontology-based models for disease, community, and intervention aspects [3]. In this approach, we use an ontology to model and specify different simulation scenarios and a translator to produce an XML-based specification for the simulation engine. This approach supports flexible modeling and allows scenario developers to use ontology editors such as Protégé [4] to define different parts of the simulation model and to reuse model components by using ontologies as a library of scenario building blocks.

Although this method provides flexibility and a systematic way of developing models, it is still limited to the expressibility of the ontology language and the subsequent translation process to the specification for the simulator engine. There are still qualities that cannot be modeled easily using this approach. For example, it is not possible to extend the capabilities of the simulation engine and introduce new features, such as intervention types and new types of data collection from the simulation process.

The aim of this study is to introduce an architectural approach to provide an additional layer of flexibility to simulations of influenza pandemics while maintaining run-time performance. Aspect-oriented programming [5,6] offers new ways of developing simulation architectures that alleviates design compromises originating from the tradeoff between run-time performance and modeling flexibility. This architectural approach allows us to add new features to the simulator without explicitly changing the core functionality and potentially degrading performance by
introducing more overhead (e.g., overhead from extension programming interfaces). Another advantage of this approach is that it is possible to use aspects in different combinations to achieve alternative configurations of the simulator. This technique makes it possible to build different versions of the simulator for specific tasks, such as additional logging, output in new formats, and new types of interventions.

**Background**

Aspect-oriented programming (AOP) is a programming paradigm that builds on object-oriented programming to provide mechanisms for modeling crosscutting concerns at the source-code level throughout the program. Here, crosscutting concerns are widespread functionality that cannot easily be modeled in a specific part of the program (e.g., as an object). Examples of crosscutting concerns are logging, consistency checks, and error handling. There are several programming languages for AOP, for example AspectJ [5] and AspectC++ [6]. These languages extend the existing programming languages Java and C/C++, respectively, with language elements for AOP rather than providing completely new languages. In AspectC++, programmers write aspect definitions in special source-code header files, which the AspectC++ compiler ac++ then transforms, or weaves, intro C++ source code suitable for compilation to executables using standard C++ compilers, such as GNU g++ and Microsoft VisualC++.

There are different modes for the weaving process. Aspect J has some additional features in that it supports weaving of compiled code (Java byte code), and allows for load-time weaving, which means that the run-time system weaves classes as they are loaded into the Java virtual machine. This type of flexibility is not available in AspectC++, because it is restricted to source-code weaving only. However, for run-time performance, there are clear advantages of source-code weaving. For example, source-code weaving (1) enables the compiler to perform optimizations on the complete (weaved) code, (2) allows aspects to introduce source-level constructs (potentially compiler directives and #pragma statements), and (3) facilitates manual inspection of weaved code if required.

One of the advantages of AOP is that it is possible to weave the program with different aspects to achieve radically different behavior from the program depending on the aspect configuration. For example, without any aspects a simulator could perform a basis simulation of an outbreak. However, with additional aspects weaved into the simulator code the program could collect additional information and statistics about the simulated outbreak, such as the gender and age distributions of infected persons. By combining different aspects, it is possible to achieve different program configurations and to develop special-purpose versions of simulators for specific tasks, such as experiments setups. In AspectC++, this type of configuration means that the program is weaved and compiled with a set of enabled aspects to form a new executable for each configuration.

**Methods**

The starting point for this work was an outbreak-simulator architecture developed in C++ [3]. As an initial step, we refactored the simulator source code to improve its structure and added unit tests (using CppUnit) to confirm proper functionality from the start. Next, we added testing aspects to ensure correct behavior from weaved code when running simulations. We then proceeded to evaluating the aspect-oriented architecture by implementing aspects for additional functionality, such as saving results directly to databases, checking consistency and supporting interventions.

**Results**

*Fundamental simulation architecture*

The fundamental simulation architecture is based on a simulator part that takes as input an XML-based specification of the simulation parameters and produces an XML file with the results [3]. In addition to the simulator, the architecture takes advantage of a scenario ontology that contains models of the community, disease, and basic interventions. Scenario developers use the ontology editor Protégé [4] to develop the scenario by instantiating concepts in the ontology. A Protégé extension transforms this ontology to the XML-based specification for the simulator core.
Figure 1 shows an overview of the architecture. The simulator consists of a simulator core algorithm with supporting modules for parsing specifications, applying interventions, logging, preparing result output, and so on. The architecture is based on object-oriented programming, which models essential modules and data components as C++ objects for efficient computations. From a user perspective, analysts can use the simulator architecture to prepare scenarios in the ontology editor, run them in the simulator, analyze the results, and, if required, revise the scenarios.

![Diagram](image-url)

Figure 1. Overall simulation architecture. The simulator takes as input a scenario ontology consisting of community, disease, and intervention models. The scenario ontology contains the necessary simulation parameters. The simulator consists of several modules with specific functions. The output of a simulation run is a compressed XML document, which contains a trace of the simulation with daily outbreak reports suitable for further analysis and visualization. End users (e.g., public-health analysts) parameterize the simulator using the ontology and control the simulator through a command-line interface.

The main advantage to using ontologies for specification of scenarios and simulation parameters is that it allows for (1) modeling flexibility supported by the expressivity of the ontology and the high-level editing tool and (2) reuse of model components across scenarios where the ontology acts as a scenario-component library. Although such ontology-based specifications work well, there are sometimes scenario properties that cannot be modeled easily using ontologies. These situations require additional flexibility. For example, the scenario developer may want to add concepts to the ontology that currently have no mapping to the XML-based parameters that simulator uses as input. Furthermore, the simulator implementation may not support these concepts at all. There is a need for a modeling middle ground where developers can add functionality to the simulator without actually interfering with the simulator code and modifying it manually (which is sometimes difficult and time-consuming to do and potentially error prone).

Aspect-oriented simulation architecture

Based on the fundamental simulation architecture, we have developed a new model for aspect-oriented simulation architectures. By taking advantage of aspects it is possible to develop simulators that are flexible in the sense that the basic simulation system can remain small and relatively unchanged while developers can add new aspects independently to change the simulator behavior significantly. Furthermore, it is possible to use the aspects in different configurations to effectively create diverse versions of the simulator architecture. As simulation
requirements evolve and usage patterns change, developers can define additional aspects and add them to the architecture.

Figure 2 shows the simulator architecture populated with aspects. In this architectural approach, the aspects add to the existing simulator code by modifying the behavior of the simulator modules at specific points and by possibly extending the data model by introduction of new member field and functions to existing classes. It is possible to define aspects that address crosscutting concerns by extending modules such as the user interface, parameter parser, data model, and output with additional functionality. For example, the developer may want to extend the simulator with functionality for taking advantage of new data sources (e.g., new types of community-related data) by enhancing the parser and other data-input modules in combination with introducing new properties to the data model.

![Simulator architecture with aspects](image)

**Figure 2.** Simulator architecture with aspects. Aspects weaved into the original simulator use pointcut expressions to add advice to simulator modules to extend their functionality. In this configuration, the aspects extend the simulator with functions for new interventions, output formats, logging, consistency checks, and so on.

**Potential simulator-architecture growth points**

Growth points are areas where it is especially interesting to extend existing systems. Developers of APIs often preplan software extensions, sometimes tacitly, by identifying areas where other developers may want to add functionality to a system. For instance, an API may support the growth point of adding import filters for new data and media formats (i.e., by providing appropriate calls and hooks). In outbreak simulators, it possible to identify candidate growth points where it is especially interesting to introduce extensions beyond what is possible to express using simulation parameters and configuration files. Note that both modeling interests (e.g., refined disease and community models) and implementation requirements (e.g., internal consistency checking and logging) can motivate such simulator growth points. Let us discuss potential simulator-architecture growth points that are possible to approach using aspects.

1. **Data sources.** New types of data as well as old data in new formats can become available to the simulator. It is possible to use aspects to develop import filters and data converter that parse and translate the input data to the internal data model. For example, aspects can add advice to an input XML parser to enable it to recognize new tags for new data elements.

2. **Data models.** The underlying data model for the simulation (e.g., classes for persons, mixing groups and other domain-specific resources) may be insufficient for certain types of scenarios. For example, public-
health professionals may want to introduce socio-economic factors in the simulation, which requires new properties in the data model. Aspects can use introductions to add to existing class definitions for central simulator concepts, such as persons and mixing groups.

3. Intervention modeling. Although it is possible to model basic interventions in the scenario ontology, there are complex interventions and combinations of interventions that cannot readily be modeled in this way. The simulator architecture can benefit from having an additional, advanced, way of modeling interventions. If needed, aspects can implement interventions using program code, which maximizes the flexibility in intervention modeling.

4. Tracing, logging, and run-time statistics. Simulators must produce sufficient records of the simulation runs, especially for lengthy jobs running for days or weeks. Developers can use aspects to implement and extend various maintenance and recordkeeping functionalities of the simulator. In AspectC++, for example pointcut expressions are useful for pattern matching of function calls for tracing and logging.

5. Consistency checks and unit tests. To ensure correct simulator operation, developers often add data-consistency checks and unit tests. It is possible to use aspects for implementing and applying such test as well as supplementing existing ones.

6. Result output. Analysts will rarely examine the raw simulation output directly. Normally, the simulation results will be stored and used as input to others tools, such as plotting and statistics packages. It is possible to use aspects to implement support for new output formats (e.g., suitable for efficient storage and processing) and to output additional data of interest from the simulation run, such as daily reports of the age of infected people.

7. Problem scalability. New scenario definitions can increase demand on the simulator, for instance in terms of community population and data granularity, which require increased memory and processing capabilities. Aspects can improve scalability by (1) changing data models (i.e., memory requirements) and (2) adding pragma statements, for example Open Multi-Processing (OpenMP) pragma directives for taking advantage of multicore processor architectures.

These growth points exemplify concerns that are fundamental to the simulator and, therefore, difficult to model as simulation parameters and in simulator configuration files. Aspects can contribute to simulator architectures by supporting such growth without developing full-grown API for each of them (which may add overhead and be difficult to implement because of their fundamental nature).

**Architecture Evaluation**

Let us consider two sample aspects as a basic evaluation of the architecture model. The aim of the first aspect is to change the output of the simulator. Figure 3 shows a sample aspect that adds an additional output format, HDF5, for the simulator results. HDF5 is a binary file format for storing large amounts of numerical data, such as measurements and other types of scientific data. This aspect adds a before advice to the simulation method to create and open the HDF5 file (by calling the H5Fcreate function in the HDF5 API) and an after advice to close the HDF5 file after finishing the simulation (by calling the H5Fclose function). Furthermore, the aspect adds an after advice to write the dataset to the HDF5 file (by calling the H5LTmake_dataset function) after the simulator has produced an array that summarizes the number of active infections for each day. The aspect defined in Figure 3 illustrates that it is possible to extend simulator functionality with a fairly compact aspect to achieve useful functionality. Moreover, it is possible to do so without requiring a formal extension API, which may unavailable or at least time-consuming to implement especially for specific situations.
Another application area for aspects in simulation architectures is implementation of interventions. Because of the variation of possible interventions, it is difficult to model many of them as simple statement or expression (e.g., an object with properties and logical expression). Potential interventions range from different types of social-distancing measures, such as avoiding physical contacts, canceling events, and closing locations, to immunization programs. Furthermore, the interventions can be applied at different times, according to different criteria, and in various combinations. Also, intervention compliance may vary. Therefore, simulation architectures require mechanisms for defining intervention models that are both simple and sufficiently expressive without actually making extensive modifications to the simulator source code.

The aim of the second aspect used for evaluation is to explore intervention modeling. Figure 4 illustrates how it is possible to implement interventions in the simulation architecture. This aspect adds functionality for closing and reopening schools in a community at different times in the simulation. The principal idea is to implement school closure by modifying the behavior of the simulator in such a way that school children do not get infected at schools. This task is accomplished by capturing the call to the method calculating the probability of people not getting infected and returning 1.0 if the mixing group is marked as closed. To control the closure and reopening of school mixing groups, the aspect adds a check before the call to the method responsible for stepping the simulation one day. This advice marks all school mixing groups as closed if the current simulation time is equal to the specified day of closure. Likewise, the advice marks all mixing groups as open once the current simulation time reaches the day of reopening. Furthermore, the aspect introduces a slice, which adds a closed property (data member), to the to the class representing mixing groups (Group). In addition, this slice enhances the class with member functions for closing, opening, and checking the closure status of the mixing group.

```java
aspect AspectHDF5 {
    hid_t file_id;

    // Create and open HDF5 file
    advice call("% Job::RunOutbreakSimulation(...)") : before() {
        std::string dbPath = tjp->target()->basePath + "_out.h5";
        file_id = H5Fcreate (dbPath.c_str(), H5F_ACC_TRUNC, H5P_DEFAULT, H5P_DEFAULT);
    }

    // Add data set for the number of infected people for each day to HDF5 file
    advice call("% Job::ReportOutbreakResultsSummary(...)") && args(summary, days) : after(unsigned int** summary, int days) {
        hsize_t dims[2] = {tjp->target()->iterationUpperBound -
                          tjp->target()->iterationLowerBound + 1, days};
        H5LTmake_dataset(file_id,"/summary", 2, dims, H5T_NATIVE_INT, *summary);
    }

    // Close HDF5 file
    advice call("% Job::RunOutbreakSimulation(...)") : after() {
        H5Fclose (file_id);
    }
}
```

**Figure 3.** Sample aspect for simulation output in HDF5 format. This aspect definition contains three advice statements for (1) creating and opening the HDF5 output file, (2) saving a summary array with the data about the number of active infections for each day, and (3) closing the HDF5 output after the simulation has completed.
To illustrate the effect of using the school-closure aspect, let us examine the effect of applying the intervention in different configurations. Figure 5 shows the results from running the simulator with the aspect in different modes and rerunning the simulations ten times for each case. The baseline case is a simulation of an outbreak in a community without any intervention (Figure 5a). The intervention aspect was then configured to close all schools on day 10 of the outbreak and reopen them on day 20 (Figure 5b) and, finally, to close on day 20 and reopen on day 30 (Figure 5c). The outbreak curves show the distinct differences in the result profiles of the configurations. Note that this experiment is an evaluation of the intervention modeled by an aspect rather than a simulation aimed at exploring

```c++
aspect AspectSchoolClosure {
    const int close = 10, open = 20;

    slice class SL {
        bool closed;
        void Close() { closed = true; }
        void Open() { closed = false; }
        bool IsClosed() { return closed; }
    };

    advice "Group": slice SL; // Introduce slice to the class Group

    advice construction("Group") && that(g) : after(Group& g) {
        g.closed = false;
    }

    // If the mixing group is closed, set the probability of people NOT getting infected to 1.0
    advice call("double Group::ProbabilityOfEscape(...)" && args(p) : around(Person const& p) {
        Group& g = *tjp->target();
        if (g.IsClosed()) {
            *tjp->result() = 1.0;
        } else tjp->proceed();
    });

    // Check if open or close conditions are met on each day in the simulation
    advice call("% Job::StepOutbreak(...)") && args(t, iter, iteration) :
        before(unsigned int t, unsigned int iter, unsigned int iteration) {
            JobSpecification &JS = tjp->target()->JS;
            if (t == close) {
                // Close all schools
                for (GroupListIterator giter = JS.Groups().begin(); giter != JS.Groups().end(); giter++) {
                    Group& g = **giter;
                    if (g.IsSchool()) g.Close();
                }
            } else if (t == open) {
                // Reopen all schools
                for (GroupListIterator giter = JS.Groups().begin(); giter != JS.Groups().end(); giter++)
                    (*giter)->Open();
            }
        }
};
```

Figure 4. Sample aspect for intervention modeling. This aspect implements school closure and reopening by (1) using introduction to add the closed boolean property to the Group mixing-group class, (2) effectively closing mixing groups by setting the probability of not getting infected to 1.0 if the closed property is true, and (3) setting the closed property to true on the day of closure and false on the day of reopening.
public health policies (which would have required additional scenarios, background data, repeated simulation runs, and result analysis).

Figure 5. Results from outbreak simulation using the school-closure intervention aspect. Each simulation is repeated ten times and the resulting curves show the number of active infections over time. (a) No intervention. Schools are open as usual. (b) All schools close on day 10 of the outbreak and reopen on day 20. (c) All schools close on day 20 of the outbreak and reopen on day 30.

Architectural model and development approach

Ultimately, aim of this work is a generalized architectural model that directs the implementation of flexible simulators. Figure 6 shows the conceptual model for the aspect-oriented simulation architecture. The guiding architectural principle is that there should be a core simulation system and data model with basic capabilities, and that aspects should add functionality in different configurations. Normally, the core simulation system is implemented as a conventional object-oriented program whereas the adjacent aspects are implemented in a corresponding aspect-oriented language (e.g., AspectC++ in the case of C++ and AspectJ in the case of Java). Using this approach, developers can implement the simulation architecture by establishing the core functionality and then add aspects around the core incrementally during the development process. For preexisting simulator implementations, developers can use a refactoring approach where they identify the core simulator part and then gradually extract existing features from the simulator system and refactor them as aspects.

It is possible to prepare the core simulation system for aspects extensions. Although aspect-oriented programming does not require an API for making extensions, it is still useful to define areas for possible extensions by paying attention to the software design. For example, it is advantageous to employ a (class and function) naming convention that enables aspect pointcut expressions to pattern match on the names. Furthermore, it is useful to promote growth points by adding (perhaps dummy) function calls at specific locations of interest in the algorithm code to allow for aspects to add advice. For example, an algorithm for calculating probabilities of infections among individuals in a mixing group can call additional (and perhaps dummy) functions at different computational stages to permit aspects to add advice through pointcut expressions.
Data model
Core simulation system
Extended parameters
Statistics
User interface
Output format

Intervention #1
Intervention #2
Intervention #3

Data source

Figure 6. Conceptual model for simulator development. The core simulation system is a small but still independently-functional conventional program that can run basic simulations. Aspects add to this functionality in different configurations.

Discussion

Aspect orientation enables flexible simulator architectures that support scenario modeling beyond what is possible in simulation parameters and configuration files. The use of aspects to extend a core simulation system has the advantage of allowing developers to improve the architecture incrementally by adding aspects individually and independently. Furthermore, this type of architecture support extensions without predefined API mechanisms, which may add run-time performance overhead especially for lengthy simulation jobs. Thereby, aspect-oriented simulator architectures provide significant advantages to pandemic influenza modelers when compared to present static simulation frameworks [7,8].

Although aspect-oriented programming is a powerful technique, there are certain modeling drawbacks. In particular, aspect-oriented code is difficult to read and debug in the sense that is difficult to know exactly what code is executing at specific locations. This is because several aspects may change and add to the behavior of functions through advice code hidden from immediate view. Nevertheless, there are jointpoint visualization tools that aid developers in reading aspect and advice code, for example the AspectC++ Development Tools (ACDT) and AspectJ Development Tools (AJDT) for the Eclipse development environment.

Developers constantly make different kinds of modeling decisions. In this type of architecture, one particular consideration is the choice between objects and aspects; that is, which modeling paradigm to follow. In some situations, it is possible to implement the same functionality as objects in the simulator core or, alternatively, as aspects added to the architecture. Our preferred guideline for such modeling decisions is to (1) use object-oriented programming for basic and common functionality, such as the simulator core, and (2) reserve aspects for extended and independent functionality while avoiding aspect interdependencies where aspects rely in the functionality of other aspects. In other words, we refrain from using aspects in the simulation core to promote simplicity and to avoid potential programming errors that could affect the simulation outcome.

An alternative to aspect-oriented programming is to use traditional object-oriented programming and define conventional APIs for extensions. This approach has the advantage of not bringing in the complexity of aspect-oriented programming and the weaving process. This method also forces simulator developers to identify growth points to support and implement well-defined interfaces among the architectural components. However, the disadvantage of this API approach is that it is less flexible than aspect-oriented programming and that it is difficult to achieve the same run-time performance as aspects because of the overhead associated with programming interfaces.
In aspect-oriented programming, it is possible to use introductions to extend the basic data model, as discussed previously. We have seen that aspect developers can extend the definition of a person (i.e., the C++ class representing persons in the simulation) with additional fields to represent properties such as age and primary-care provider. Doing so in programming interfaces is possible, but at the cost of additional classes and computational overhead. For example, object-oriented APIs often take advantage of polymorphism and virtual member functions (methods) to allow extension by addition of new subclasses. Unfortunately, the use of virtual function calls in tight simulation loops can be quite costly (because of the dynamic run-time lookup of virtual member functions). Aspects can avoid such costly look-up operations by weaving advice with regular member functions.

Conclusion

Simulation of influenza pandemics is important for response planning because of the possibility to predict outbreak dynamics and the relative benefit of interventions before empirical data are available. In general, there is a trade-off between flexibility in configuring simulators and run-time simulator performance. In this work, we have explored a novel simulator-architecture approach based on aspect-oriented programming. The results illustrates that it is possible to develop simulator environments based on a core system combined with extensions implemented by aspects. The aspect-oriented approach makes it possible to avoid the potential overhead of using programming interfaces and generalized system designs with complex mechanisms for extensions. In summary, we believe that this type of flexible and extendable architecture is an important step towards rapid modeling and simulation for pandemic preparedness and response.

Acknowledgements

The Swedish Research Council supported this work under contracts 2008-5252 and 2009-6291.

References

Designing Asynchronous Communication Tools for Optimization of Patient-Clinician Coordination

Jordan Eschler¹, Leslie S. Liu¹, Lisa M. Vizer, PhD¹, Jennifer B. McClure, PhD², Paula Lozano, MD, MPH³, Wanda Pratt, PhD¹, James D. Ralston, MD, MPH²
¹University of Washington, Seattle, WA; ²Group Health Research Institute, Seattle, WA

Abstract
Asynchronous communication outside the clinical setting has both enriched and complicated patient-clinician interactions. Many patients can now interact with a patient portal 24 hours a day, asking questions of their clinicians via secure message, checking lab results, ordering medication refills, or making appointments. However, the mode of communication (asynchronous) and the nature of the interaction (lacking tone or body language) strip valuable information from each side of patient-clinician asynchronous communication. Using interviews with 34 individuals who actively manage a chronic illness of their own, or for a child or partner, we elicited narratives about patients’ experiences and expectations for using asynchronous communication to address medical issues with their clinicians. Based on these perspectives, we present opportunities for designing asynchronous communication tools to better facilitate understanding of and coordination around care activities between patients and clinicians.

Introduction
Asynchronous communication between patients or caregivers and members of their care team (physicians, nurses, and medical assistants, herein referred to as “clinicians”) has increased with the adoption of patient websites or portals. Patients may now interact with these electronic systems to manage basic health care needs, including filling prescriptions, receiving test results, being reminded of future screening and diagnostic testing, and sending secure communications to health-care providers. Some of these services are also now part of Federal policy. To receive full payment for seeing patients with Medicare and Medicaid coverage, Stage 2 Federal meaningful use criteria for electronic health records (EHRs) require healthcare providers to support patient use of secure electronic messaging and the ability to view, download and transmit portions of the electronic medical record. Several studies have shown the potential of these services to improve patient satisfaction and outcomes in patients with selected chronic conditions.

Despite the growing role of asynchronous communication between patients and healthcare providers via portals, patients have had limited involvement in their design and implementation. Asynchronous communication increasingly extends beyond secure electronic messaging and now includes reminders for preventive and chronic care tasks, notifications about medical test results, and patient uploads of survey and biometric data like home blood glucose and blood pressure readings. Outside of a few core features associated with high patient satisfaction, we understand little about what portal features can support the needs of patients for asynchronous communication and information, particularly for those living with chronic illnesses. Several underserved patient populations—including those from racial and ethnic minority backgrounds, and those with lower levels of education—are also less likely to use patient portals, raising concern that current portal design and implementations are not equitably accessible to those in need. Finally, asynchronous communication over portals is largely limited by what is provided by commercial EHRs and the possibilities of integrating with existing models of care.

To improve asynchronous communication over patient portals, we must consider the patient’s perspective of their use, effectiveness, and usability. The design of asynchronous communication should align with the Epstein et al. vision for patient-centered care, where care should “explore patients’ values and preferences; help patients and their families make clinical decisions; facilitate access to appropriate care; and enable patients to follow through with often difficult behavioral changes needed to maintain or improve health.” To this end, we elicit experiences from patients and caregivers in their use of patient portal asynchronous communication tools. To gather perspectives of patients who have ongoing health needs and are most likely to benefit from asynchronous communication, we recruited adult patients with diabetes and mothers of children with asthma. We oversampled racial and ethnic minorities and ensured an adequate sample of those with lower educational status to identify potential portal features that would engage these populations in asynchronous communication. Our study contributions are twofold: first, we present themes in patient experiences of asynchronous communication tools. Then, we suggest design implications
for asynchronous communication tools to better facilitate coordination of care between patients and clinicians between office visits.

**Background**

We frame our approach to eliciting patient experiences in using asynchronous communication tools as just one factor in chronic illness management and in the larger relationship with clinicians to support such management. Similarly, Abaidoo and Larweb situate information and communication technology in clinical practice as “[a] conduit for channeling health information to consumers” for two purposes: (1) making informed shared decisions, and (2) tailoring clinical interventions appropriately through timely communication. This definition conveys agency in decision-making to both patient and clinician, underlining the importance of understanding care activities and their priority in the patient-clinician relationship. Thus, communication facilitated by technology is one factor that can help a clinician and her patient negotiate goals, establish benchmarks of wellness, and manage health and/or chronic conditions. Patient access to medical information and communication tools is, accordingly, an important tenet of the Institute of Medicine’s recommendations to “cross the quality chasm” in health care.

Fowles et al. found that 80% of patients—regardless of health status, education, and income—had at least some interest in viewing the information in their medical record; respondents to the survey in this study were motivated by wanting to be more involved in their health care and by a desire to understand their medical conditions more thoroughly. Therefore, patients do desire to engage with their health information through tools such as patient portals. These tools have the potential to increase patients’ engagement in their own care, improve patient experience of care and, potentially, improve patient outcomes. A systematic review by Goldzweig et al. found that patients had generally positive attitudes towards secure messaging and access to the online medical record. The authors further found improved outcomes for certain chronic conditions when secure messaging and other functions of the portal were coupled with case management. However, the review also identified the importance of further evidence on the specifics of context and implementation in evaluations of portal functions. Much of the evidence of for patient benefit has come from implementations and studies in a limited number of settings.

Clinic-focused evaluations of asynchronous messaging systems assess clinician engagement, associated workload, and billing. In a meta-review of such evaluations, Wallwiener et al. found that physicians were slow to take up using secure messaging, but found benefits—such as less time spent on the phone and higher measured patient satisfaction—outweighed costs, particularly burdens on physicians’ clinic time (also supported by Kummervold and Johnsen). Goldzweig et al., however, found insufficient evidence that patient portals change costs or traditional forms of utilization. Tufano et al. highlighted the importance of aligning the organization and financing of care for providers to sustainably meet patients’ needs for access over secure messaging.

Benefits of asynchronous communication tools to both patients and patient-clinician relationships are multiple. Andreassen et al. found that benefits to patients included: “emancipation” of the patient-clinician relationship from the constraints of space and time during office visits; a valuable mechanism for “transferring responsibility” to the doctor for an issue once communicated; and a modality (writing) that helps patients to reflect and synthesize health information. Another benefit of asynchronous messaging cited by Sun et al. is the allowance for patients and doctors to communicate directly. As a result, the authors argue, “accuracy and authenticity” of patient messages was increased, a worthwhile tradeoff for the added time on the clinicians’ part spent addressing patient messages.

However, individual patient experience with asynchronous communication tools have principally been studied by quantifying engagement with and use of the system through patient message volume, and surveys of patients to examine use, expectations, and experience. Our qualitative study examines asynchronous communication tools as one aspect of all elements of care in managing chronic illness, a process that involves clinicians but takes place largely in the patients’ home and between office visits.

**Methods**

To inform our methods, we referred to recent workshop reports in the areas of human-computer interaction and medical informatics, which call for framing inquiry into patient-clinician communication from a patient-centered point of view. That is, we are interested in the patient experience of “asynchronous, remote” modes of communication in the context of the larger patient-clinician relationship. In shaping our interviews with patients and caregivers, we also emphasized “shared decision-making” between patients/caregivers and clinicians. In sum, we sought to answer:
How do individuals characterize their experiences of and expectations for using asynchronous communication strategies to coordinate health care with clinicians?

We recruited 34 participants from Group Health Cooperative, a large integrated healthcare delivery system in Washington State. All participants had a primary care provider in a Group Health owned and operated clinic. Participants included 16 mothers with children under 12 years of age who were diagnosed with asthma and being treated for asthma, and 18 adults diagnosed with Type 2 diabetes. Mothers of children with diabetes (mean age 38.3) were younger than patients with diabetes (mean age 73.0; also see Table 1). We oversampled racial and ethnic minorities (16 of 34 enrollees, the greatest proportion of whom were Black participants). Nine participants had a high school or lower educational level. We purposively sampled 26 (76%) of the overall interview sample to be current users of the patient portal. We defined portal use as having used one or more of the eight key services of the portal on two occasions separated by 30 days or more in the previous two years. These services included: viewing medical test results, visit summaries, immunization lists, allergy lists, medical condition lists; exchanging secure messaging with providers; ordering medication refills; and scheduling an in-person appointment.

In the case of diabetes patients with a close caregiver, the caregiver participated in the research interviews where possible and appropriate. These two cohorts were chosen because each is responsible for daily self-care related to the management of a chronic disease, each has frequent contact with health care providers and health care systems (e.g., scheduling appointments, filling prescriptions, scheduling lab tests, etc.), and all have regular opportunity and need to use communicate with their healthcare providers. Asthma cohort participants are denoted with A# identifiers and diabetes cohort participants have D# identifiers in the quotes highlighted in results.

<table>
<thead>
<tr>
<th></th>
<th>Asthma cohort</th>
<th>Diabetes cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Male</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>Average Age (years)</td>
<td>38.3</td>
<td>73.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education</th>
<th>Asthma cohort</th>
<th>Diabetes cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to HS/GED</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Some College/AA</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4 Year Degree</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>More than 4 year degree</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race &amp; ethnicity</th>
<th>Asthma cohort</th>
<th>Diabetes cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Black</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>White (Non-Hispanic)</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>White (No ethnicity given)</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Other (Hispanic)</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

We conducted 2 interviews each with all of the 34 individuals who were each managing a chronic condition of their own or for a young child. Interviews took place approximately six months apart, were conducted in the participants’ homes, and ranged in length from 30 minutes to 90 minutes. The semi-structured interviews inquired about patients’ health goals, priorities for completing health tasks, and examined patient-centered workflow in attaining health goals and completing related tasks. The research design was approved by institutional IRB, and researchers used a semi-structured script to guide the sessions in both interview instances. Interviews were audio recorded; after transcription, audio files were destroyed. We had approximately 1,300 pages of interview transcriptions as our dataset for this study.
Analysis
The group of researchers conducting interviews and those coding the transcripts overlapped; thus, as interviews were being conducted - particularly in the second round - researchers tracked themes emerging in real time (as described in the qualitative research approach by Strauss and Corbin\textsuperscript{24}). One author (JE) performed open coding on six pairs of interviews, using a list of potential themes identified during the interview period to seek patterns and name and define potential codes. Three other coders (LMV, LSL, KO) were trained with the codebook and tested its use on four pairs of interviews each. The four coders then met to revise and finalize the codebook for its use in pursuing themes to address the research question.

Given the variation in our populations of interest and their technology use, we were careful to inquire about multiple possible modalities of communication between patient and clinician (not only asynchronous), to obtain a full picture of the patient experience. Therefore, we were able to assess themes that emerged around the patient-clinician relationship, as well as patient motivations for utilizing their modalities of choice in coordinating with health care providers. We examine below two characterizations of patient experiences: themes that emerge in satisfactory patient-clinician asynchronous communication and those that coincide with unsatisfactory communication.

Results: Characteristics of satisfactory asynchronous communication
We grouped themes of asynchronous communication tool use that were consistent with positive or satisfactory patient experiences. Themes in satisfactory communication experiences included experiences where asynchronous communication enhances care, reduces patient uncertainty, and provides “health archives.”

Enhances care with follow-up
Participants contextualized their asynchronous messaging experiences in the relationship with their doctors. D07 described his doctor as someone who followed up diligently: “I had expressed a concern to her and she checked into it and wanted to know if I was still having a problem and she e-mailed me and I e-mailed her back.” Another participant used e-mail as a trusted communication medium because the doctor was skillful in judging when to follow up. “Some things require a response and others are just for their information, it depends.” (D03).

Participants felt that asynchronous communication was particularly beneficial when the communication complemented care received during a clinic visit, particularly when communicating between clinic visits also elevated the sense of patients’ responsibility for managing their conditions. For example, D09 expressed that “if you don’t answer [clinician e-mails], that would be inappropriate and certainly not taking care of yourself.”

The additional modality offered by asynchronous communication (i.e., writing) also assisted patients in following up more effectively. Participants who were satisfied with the asynchronous communication system tended to indicate comfort with expressing themselves in writing; this echoes previous research findings\textsuperscript{25,26}. D15 stated, “I can sit down and express myself more.” One participant in the diabetes cohort felt less pressure in office visits to cover all concerns, and felt more control over her health when able to follow up post-visit:

I’ve already left the doctor, I’m feeling a certain kind of way. I don’t know that my inadequate feeling is enough to make another appointment so it depends what it is that I’m feeling and then I can go ahead if I need to contact them, to do the email [secure messaging]. I’m just so pleased they got the email thing. (D03)

Asynchronous communication tools were also key to managing complex family arrangements, such as when non-custodial parents brought children to the doctor for a checkup. Asynchronous communication tools were used, in following up clinic care, to keep the custodial parent “in the loop.” In one case, a custodial parent used secure messaging to (1) tell the doctor ahead of time that the other parent would be accompanying the child, (2) send along health concerns in her absence, and (3) follow up with questions between custodial and non-custodial parents after the appointment. This participant, A09, explained: “It’s my control issue, this is what I need [the doctor] to cover, or could you just answer this for me or know this information when you talk [to the other parent].”

Reduces uncertainty in the plan of care
Participants also used e-mail to resolve uncertainty in the plan of care, such as questions that remained after an office visit. In some cases, the participants felt as though they needed less assistance from clinicians because they could access archived reports (such as after-visit summaries) and messages through the patient portal: “When I feel I need follow up I can check there, which is really nice that they have the online summaries.” (A16)
One participant with diabetes frequently used the patient portal’s secure messaging function to discuss his lab test results with clinicians. This not only reduced uncertainty, but increased his feeling of “buy-in” on engagement in his care. He felt his clinicians emphasized their responsibility to make sure the patient was adherent by using secure messaging frequently: “I don’t get by with ignoring anything,” he told us (D01). In another example, one patient on Warfarin had frequent contact with consulting nurses and indicated that the secure messaging was both informative and an indicator of the seriousness with which the participant should be managing his health: “I’m very happy that they follow up so diligently, because previously I didn’t think [the Warfarin monitoring] was that serious a thing, really” (D09).

Provision of automatic “health archives”
Participants also acknowledged that using asynchronous communication systems, such as secure messaging, had the advantage of automatically archiving their communication with clinicians, which they could review at a later date to inform their care management. Like many participants, A06 told us, “I don’t save paper notifications, but everything else is stored online.” The online archive of messages can also help patients understand their condition better. Caregivers also used the e-mail archives to access information as part of their responsibilities. One caregiver of a partner with diabetes (D10) told us, “Sometimes I check his [patient portal messages] because I know he doesn’t do it.” In this case, the archive of information facilitated social sharing of information that helps the patient manage his condition.

**Results: Characteristics of unsatisfactory communication**
The themes below were identified as characterizing unsatisfactory patient experiences with asynchronous communication tools on the portal, and included: failing to track issues in a coherent way, and exposing patients to inconsistent communication patterns.

**Fails to track issues**
Participants indicated that they had, in some cases, been “conditioned” by the system about expectations, such as response time on secure messages or lab test results. “I just go into [the patient portal] when I expect [lab results] should be done, and check it and see if it’s done” (D07). However, some of the conditioning by the system trained participants to believe that “no news is good news,” and if they never heard back from the doctor, the issue was not only no longer open, but in fact a non-issue. As one diabetes caregiver to participant D16 explained, “Unless there’s something really out of whack, and then [the doctor’s] office will call…[but my husband] is actually really well stabilized.” The uncertainty in these cases led participants to assume the burden was on clinicians to follow up as needed, but this is not ideal for optimum understanding of and coordination around illness management activities.

Participants also cited the inability to act on reminders to revisit an issue after a certain period of time. For example, a mother of a child with asthma is often advised to visit the child’s allergy specialist in 6-month intervals, but the mother is unable to schedule that far ahead on the patient portal system. Instead, she tries to call for an appointment immediately; if that tactic fails, she implements a reminder-to-remind system:

*If I go in for an appointment and the doctor says, ‘okay, we need to see [the children] back in six months,’ I’ll usually get home, I’ll call the next day, even though it’s six months out – sometimes their calendars don’t even, they’re not even that far out. But I call the next day because I might forget so well, I’ll just call and say, ‘okay, are you scheduling this far out?’* (A07)

A related problem was the lack of status indicators for unresolved issues, where communication tools on the patient portal were either “read” or “unread,” or had no useful status to keep participants informed of ongoing tasks associated with open issues. A15 wished for open medical orders and pending lab results to be flagged as open issues on the patient portal; she stated that such status information would give her “peace of mind.”

Secure messaging can be part of overall unsatisfactory communication leading to severed relationships between the patient or caregiver and clinicians. In one such episode, the mother of a child with asthma and a comorbid sleep disorder described a lack of follow-up from a provider on an expensive sleep study. “They did ask, do you prefer [a secure message] or [a] phone call? I said ‘either one,’ and we never received anything. And in turn, with that specialty...we have gone out of network.” (A04) Had the patient portal offered a way to track the status of issues associated with the sleep study, the clinician would have been notified that the issue remained open for the caregivers and patient.
Exposes Patients to Inconsistent Communication Patterns

Some participants were confused about how asynchronous communication tools in the patient portal worked within the larger health care information system. The patient portal, along with the multiple modalities of communication available to clinicians, formed a “black box” of a system where patients were not always certain what to expect. One patient never knew how she would hear about test results: “I think my mammogram, they never called me for that - they do send you a letter, so they aren’t consistent” (D01).

In a similar vein, another patient noted that although medical orders are being executed, there is no communication with the patient to ensure an order is completed: “They put [the lab test order] on the computer, but they don’t tell you to go in and have it done” (D13). Uncertainty in communication even affected the frequency of in-person office visits, confusing the patient and leading to potential lapses in illness management: “I haven’t seen [doctor] for six months. I don’t know when I’m going to see her again...Maybe I don’t need to, I don’t know.” (D13)

In the following section, we synthesize what we have learned about successful and unsuccessful patient experiences with asynchronous communication tools. We suggest design opportunities for such tools that could improve the patient experience, and increase patient-clinician understanding of care activities to manage chronic illness.

Discussion of Design Opportunities

Based on our findings, we have identified two major design opportunities for enhancing patient-clinician coordination when using asynchronous communication systems. First, incorporating patient preferences for communication modality in a transparent manner within the patient portal would facilitate a greater match between clinician efforts to remind and patient ability to receive and act on reminders. This is relevant to supporting Meaningful Use Stage 2 guidelines\(^2\), which include patient communication preference incorporation to improve patient access to information and online health content. Second, incorporating “status indicators” in the asynchronous communication system, giving both the patient and clinician the ability to indicate a range of statuses for individual care issues, such as “unresolved,” “under clinician review,” “awaiting clinician response,” and “resolved,” would give patients and clinicians a signaling system for prioritizing future tasks and indicating agreement on clinical tasks and issues.

To demonstrate that implementation of functions related to these design opportunities would improve patient-clinician and caregiver-clinician communication and coordination, we use the model proposed by Wilcox et al.\(^24\) to delineate benefits to care. Specifically, Wilcox et al. called for research to focus on three “communication goals” between patients and clinicians: (1) clinician awareness of patient symptoms, (2) patient awareness of care activities, and (3) shared decision-making. An overview of our design recommendations and their benefits according to this framework are presented in Table 2, and we further detail design recommendations in the following subsections.

Incorporating patient preferences

We found that preferences for different modalities varied within our sample cohorts, and preferences were highly individualized. This supports the requirement that providers begin to collect patient modality preferences (e.g., secure messaging, phone) as part of Meaningful Use Stage 2\(^2\), as the act of incorporating preferences seems to be a formidable opportunity to improve the patient-clinician communication loop. For example, some of the patients with diabetes strongly preferred messaging to using the phone, due to the low time pressure of writing messages and the assumed convenience to clinicians. “On the phone, the guy's got to be there and on the e-mail, once you send it, you send it and you know he's got the message” (D08).

However, some mothers of children with asthma preferred never to check their patient portal messages and many described their difficulty with remembering yet another password for an online account. Instead, mothers who spent a great deal of time traveling in the car, for example, preferred to see the clinic main number pop up on their smartphones, which acted as a reminder in and of itself:

Interviewer: It sounds like you don't get the calls from [all of your doctors] right now. Would you prefer texts over calls in general?
Participant: I really don't have an opinion. I mean if you can get the information out in a text, why not? But if not, then a call is fine too. At least I can see the missed number and recognize the number so I mean - you know what? Even saying that out loud, it sounds like I would prefer the visual, so yeah, I would prefer text. (A09)

These variations in modality preferences for non-urgent communications indicate that customizing modalities for asynchronous communication would offer a great benefit to clinicians (saving them time on communication issues) and to patients or caregivers (reducing uncertainty around how communication with clinicians would occur). In times of greater need, such as with an asthma exacerbation or worsening complications from diabetes, the benefit may be amplified further.

Table 2. Delineating benefits to design opportunities identified in the Wilcox et al.\textsuperscript{24} patient-centered communication framework

<table>
<thead>
<tr>
<th>Wilcox et al. Model elements</th>
<th>Design Opportunities Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Clinician awareness of patient symptoms</td>
<td><strong>Patient</strong> more likely to follow up with open issues when communication uses preferred modality. <strong>Clinicians</strong> will have better information to act upon to help patients.</td>
</tr>
<tr>
<td>(2) Patient awareness of care activities</td>
<td><strong>Patient</strong> more likely to receive and act upon reminders for care activities through preferred modality. <strong>Clinicians</strong> will benefit from improved efficiency in reminding patients about care activities.</td>
</tr>
<tr>
<td>(3) Shared decision-making</td>
<td><strong>Patients</strong> and <strong>clinicians</strong> are more likely to connect efficiently over the patient’s preferred modality to make shared decisions.</td>
</tr>
</tbody>
</table>

The reduction of uncertainty to patient expectations would in turn increase the coherence of communication through the patient portal. Rather than the confusing “black box” mental model, patients would be able to expect a phone call or secure message – depending on their preferences – and the stress of awaiting information from clinicians could be reduced, particularly around test results:

“*They send you a letter telling you what your [lab test] results were, but it's not on [the patient portal]. Like it's totally separate, and they don't do it through e-mail, which is annoying because I don't like to call people and bug them.*” (D01)

This patient with diabetes did not know what to expect when awaiting test results; as with the “black box” problem before, the confidence in knowing what to expect can be troublesome for building trust in the patient-clinician relationship. Thus, we see an opportunity to let patients determine their preferences in a way that they can see in the system, such as a dashboard to view and update preferences about communication for various tasks, like receiving test results or receiving appointment reminders.

Finally, there are situations in which asynchronous communication simply may not work well. For example, one of our participants has been visually impaired since birth. This individual stated that she preferred phone calls with clinicians to any kind of asynchronous communication, such as secure messages or U.S. Mail (which required a software screen reader, or a paid human reading assistant, to convey this sensitive information). Although an edge
case in terms of context, we point to this individual’s experience to illustrate a case in which the utility and efficacy of notifications could be improved significantly by incorporating patient preferences:

Interviewer: Have you ever considered tools that might be helpful to [remember upcoming appointments]?
Participant: I’ve asked them to call me ahead of time to remind me, and they said they don’t send reminder calls.
Interviewer: That’s surprising to hear, actually. That was something you asked the front desk about?
Participant: Yes.
Interviewer: Did you ask your doctor directly about it as well?
Participant: I think I did, I don’t recall if I did or not. (D17)

The examples above are intended to highlight the frustrations patients experience when their preferences cannot be incorporated into patient-clinician communication. This is the first design opportunity we identified in our data analysis; the second design opportunity is that of using status indicators to signal open, in-process, and closed issues between the patient and clinician.

Indicating Status
Failures and frustrations with asynchronous communication between participants and clinicians largely stemmed from the participants’ uncertainty about the status of a given issue. It is often unclear to patients whether the lack of follow-up or response from a clinician indicates “no news is good news” or simply a lack of resolution. The most radical aspect of this design recommendation is, perhaps, giving patients the ability to change and control the status indication. Both the patient and clinician would be able to view and add information to open issues. However, the patient would indicate when an issue was resolved satisfactorily.

By empowering patients to send clear signals about whether an issue is open or closed, clinicians will be better able to understand a patient’s state of mind regarding medical concerns, and can prioritize issues that patients needed further assistance to close. For example, A15 brought up an instance where she did not hear back about sensitive lab test results over a long weekend. She stated she would have felt better to have just known from the clinician “we’re looking at your test results.” In this case, the participant would have benefited from a status indicator system that would have flagged the issue as unresolved, but in-process at the clinic level, to reduce her uncertainty regarding the open issue. Such feedback has been noted from parents managing children’s health care through patient portals in previous studies, where even a “read receipt” type feedback mechanism would have been a helpful system feature in the caregiver- or patient-clinician communication loop27. Additionally, previous studies have shown that patients managing a chronic illness are willing to engage with online health management, though clear communication “closed loops” between patients and clinicians would improve the patient experience28.

Often, participants wished for a way to notify clinicians that the issue remained open, despite the clinician having moved on. D08 shared an anecdote about attempting to get referred for foot care, but described an e-mail exchange with his primary care physician that was “never successful” (i.e., resolved) - and he continued to go without foot care or a reason for no referral.

Interviewer: How can [your doctor] be there for you? Tell me more, I want to get more insight there.

Caregiver: To share a lot more knowledge and let us know why decisions are made. We've asked him – ‘oh, I'll check on it,’ but we never heard back from him on it. Every time we go there, we ask the same question – ‘why was it denied, is there something you can do?’ It doesn't seem to be one of his top priorities. (D16)

In this case, the patient would benefit from being able to indicate that an issue was not satisfactorily closed, ensuring they receive all of the information needed to move on to other care issues and reducing damage to the patient-clinician relationship.

We suggest testing the addition of status indicators to issues opened by patients or clinicians, whether it arises during an office visit and subsequently requires follow-up (lab tests), treatment (new medications, therapies), or monitoring (ongoing lab tests, such as an A1c, or home measurement of peak flow meter). Open issues could be tracked, and the corresponding information and related tasks grouped, for the convenience of both the patient and the clinician; we recommend prototype testing to maximize the benefit of status indicators and to avoid alert fatigue among either patients or clinicians. This design feature could also indicate the state of the issue, such as if the next
task required input from the patient or clinician, clarifying expectations of both parties. Having a signaling system between the patient and clinician could both reduce uncertainty and impart shared, but clear, responsibilities to each party regarding resolution of the issue.

Status indicators could assist patients in following through with standing orders for regular labs, such as the A1c for patients with diabetes. In the current system, as D01 told us, “The A1c has been a little bit of a problem because my doctor has set it up so that I can go every three months, but if that time passes...they just discard it if you don't use it.” In a future system, with status indicators for open issues, this participant could log into the patient portal and see she had an open status item to complete: she could then visit the lab, so as not to waste the standing order. This feature would underscore the importance of teamwork between patients and clinicians, and conveys a sense of responsibility to the patient in managing symptoms or problems.

Limitations
The study findings are limited by the sample size, qualitative nature of the data, and inclusion of participants from a single health care system who all have access to the same communication systems and patient portal. However, the findings suggest opportunities for improving patient-provider communication and systems diverse health care settings.

Conclusion
As evidenced by our analysis, the patient experience in using asynchronous messaging tools could be improved to enhance coordination between patients and clinicians. Our sample included individuals from populations who have been less likely to engage in patient portals, those who avoided or reluctantly used technology, as well as avid smartphone users, to give a range of viewpoints on asynchronous communication tool use in the health care context. We found that the positive aspects of patient experience with asynchronous communication tools support the ability to supplement in-clinic communication. This support included: follow-up with care activities, reduction of patient uncertainty about care plans, and availability of “health archives” online. Negative aspects of using such tools, however, included lack of consistent patterns of use in patient communication and uncertainty about resolution to issues addressed through the tools. We present design opportunities to reduce the uncertainty that patients experience in managing chronic illnesses, while also strengthening roles and responsibilities between patients and clinicians in managing issues that arise in monitoring chronic conditions. We use our findings to provide practical, actionable changes to the design of asynchronous communication tools to improve the patient experience and facilitate the balance of work and responsibility in the patient-clinician relationship.

Acknowledgements
This project was supported by grant #R01HS021590 from the Agency for Healthcare Research & Quality (AHRQ).

References

Reviewing 741 patients records in two hours with FASTVISU

Jean-Baptiste Escudié, MD1 2, Anne-Sophie Jannot, MD, PhD1 2, Eric Zapletal, PhD1, Sarah Cohen, MD1 2, Georgia Malamut, MD, PhD1, Anita Burgun, MD, PhD1 2, Bastien Rance, PhD1 2

1University Hospital Georges Pompidou (HEGP); AP-HP, Paris, France
2INSERM; UMRS1138, Paris Descartes University, Paris, France

Abstract

The secondary use of electronic health records opens up new perspectives. They provide researchers with structured data and unstructured data, including free text reports. Many applications been developed to leverage knowledge from free-text reports, but manual review of documents is still a complex process.

We developed FASTVISU a web-based application to assist clinicians in reviewing documents. We used FASTVISU to review a set of 6340 documents from 741 patients suffering from the celiac disease.

A first automated selection pruned the original set to 847 documents from 276 patients’ records. The records were reviewed by two trained physicians to identify the presence of 15 auto-immune diseases. It took respectively two hours and two hours and a half to evaluate the entire corpus. Inter-annotator agreement was high (Cohen’s kappa at 0.89).

FASTVISU is a user-friendly modular solution to validate entities extracted by NLP methods from free-text documents stored in clinical data warehouses.

Introduction

The worldwide trend toward secondary use of health data for research opens up new and exciting perspectives for researchers in all the realms of medicine. The large adoption of EHRs provides a steady flow of data, structured and unstructured (and particularly free-text medical reports). Today, almost all the preeminent clinical research institutions have deployed Clinical Data Warehouses (CDW) 1–4, to store and integrate all the data produced in routine care.

Our institution, the European Hospital Georges Pompidou (HEGP) is a 730 beds public research hospital in Paris. An i2b2 CDW has been installed and used since 2008. It integrates data from more than 700,000 patients. 5 The large majority of the data stored is structured data with more than 100 million data points of lab results, but the data warehouse also integrates more the 4 million free-text reports in French ranging from discharge summaries, letters, to imaging and pathology reports.

Structured data, by their organized nature, are a target of choice for secondary use. However, free-texts are at least equally important. In medicine, free-text has always been the support of medical records. Free-text reports collect any information that the physician considers of any importance and that took a role in a medical decision. Text reports also include data that might be difficult to find elsewhere such as family history, results of exams performed outside of the institution, rejected hypotheses. Nevertheless, the extraction and reuse of data from free-text reports requires processing and dedicated tools and resources.

Structured and unstructured data capture different views on the patient and her/his disease. Billing codes, often coded with ICD-9 CM or ICD-10 capture the medico-economical aspect of a medical encounter. The main purpose is to evaluate the cost of the treatment provided; lab results coded in LOINC provide standardized biology lab results. Issues might occur when researchers attempt to leverage structured data in a context for which they were not designed. For example billing codes are not aiming at providing a global coverage of comorbidities of a patient. In their article, Li et al. 6 compared information found in structured data to information found in free medical text. They reported two types of differences: information completeness and concept granularity, and concluded that data extracted from free-text complement structured data.

Medical documents come in a large variety of type of texts. The biomedical literature reports text mining of radiology reports7–8 medical observations 9,10, nurse narratives 11 – documents produced for care activity – but also from documentations and guidelines 12. Most of these clinical documents are stored in CDWs. However they provide information of heterogeneous quality: for example a discharge summary is validated by a senior physician while observations or notes may be written by medical students.
Standardized text annotation is required to enable automated processing. Terms in the text are annotated using standard terminologies. The Unified Medical Language System® (UMLS®) is often used as a lingua franca for biomedical domains. Tools have been developed to assist annotators, such as the BRAT Rapid Annotation Tool 13 (BRAT is not dedicated to medical applications). However, manual annotation is a tedious and time consuming process that requires to follow guidelines, and therefore to be trained. The Natural Language Processing community has developed over the last decades many applications for virtually all the fields of medicine. For example, for the detection of incidental findings in chest x-ray 14, pneumonia or rheumatoid arthritis identification from narrative reports 15–17. There is a large variety of tools available for different purposes.

Medical concept recognizers (MER) leverage medical terminologies and linguistic resources to identify medical entities in free-text (popular MERs include MetaMap 18 developed by the National Library of Medicine, the Biportal Annotator developed by the National Center for Biomedical Ontologies – NCBO, or cTakes developed by the Apache Software Foundation). In addition to recognizing medical concepts, it is often needed to capture the context of the information. The meaning of concept can drastically change depending of the context (e.g. negation, potentiality, family history). Some of the previously cited tools include some context recognition. The most commonly used context recognizer is probably NegEx 19 for negation 20 and its extension CONText 21. Expert validation is often required to secure further use of the results. I2b2 itself proposes NLP tools through its ‘Natural Language Processing cell’ based on a HITEx 22 core. This functionality is encapsulated into i2b2 interface where the user can retrieve different types of concepts detected in text: diagnoses, discharge medications, smoking status.

Shivade & al. 20 reviewed 97 articles describing approaches to identifying patient phenotype cohorts using EHRs. Forty-six articles used NLP based approach.

A common task in a clinical research is the selection of eligible patients for a clinical study. It is a tedious and costly task that requires time and highly skills personal. In the case of retrospective studies, it is often required to browse through the entire record archive of a patient to identify relevant information. Tools have been developed to assist in the selection process. They provide a connection between CDW and NLP tools, and help the user in the selection process. Many systems use selection systems inspired by the “basket” from online-shopping websites 23, i.e. the clinician browse through available documents and identify those of interest.

However, it is often needed to be, not only able to identify patients that fit a set of criteria and are eligible for a study, but also to identify and annotated the presence, or the status of phenotypes in their record. We did not find a publicly available solution that would allow a flexible entity selection, and allow phenotypic annotation by one or more users.

Moreover, most of the documents available for research in our hospital are in French. Non-English languages lack tools and terminology, and despite substantial effort toward better NLP coverage simple solution are still needed. We needed a solution with good performance, reducing physician time involved reviewing and classifying features on medical records for different tasks from patient eligibility to phenotypes extraction.

We have designed FASTVISU as an application that connects NLP outputs with free-text from CDW and provides an interface to efficiently select a set of interesting features. FASTVISU allows a single or multi-user evaluation of several features at the same time at the patient, visit or document level, based on a voting mechanism. In the next sections, we will describe the architecture of FASTVISU, and present a use case on the annotation of a cohort of patients with the Celiac Disease.

Material and Methods

Architecture

The modular FASTVISU architecture is based on a Web Service delivered as a PHP REST API. The API provides abstraction layers to:

- Retrieve documents from an i2b2 CDW;
- Create, modify and retrieve corpora of documents (stored in an Oracle database);
- Vote on features to classify patients (e.g. to select patients having a diagnosis of interest), encounters (e.g. to differentiate normal follow up encounters from relapses) or documents (e.g. classify legally conform documents from non-conform during an administrative audit). Votes are stored in an Oracle database.

554
The clients using this API are of two types:

- A user interface for human interaction (Figure 2)
- *bots* who automatically go through every document and vote according the computation from text mining modules.

![Diagram of FASTVISU architecture](image)

**Figure 1.** Modular architecture of FASTVISU.

![Screenshot of web client](image)

**Figure 2.** Screenshot of the web client visualizing a whole patient record on the left panel (using regular expressions from the text processing module to highlights concepts). The panel on the right summaries instances of concepts (color coded) and provide the user with voting options for the different categories of concepts (e.g. ‘presence of Type 1 Diabetes Mellitus: Yes/No’).
Clinical Data Warehouse

I2b2 is a widely used CDW which became de facto a standard over the years. The database of the CDW at HEGP contains data for more than 700,000 unique patients. Data available include full medical and administrative records (for in- and out-patients), diagnoses and procedures codes, EHR, structured observations, free-texts (discharges, letters), laboratory tests and pathology results, as well as every drugs prescription done in the hospital.

Evaluation

We evaluated the performance of FASTVISU to help identifying the autoimmune comorbidities associated with celiac disease in the HEGP i2b2 data warehouse. Celiac disease (CD) is an autoimmune disorder induced by the intake of proteins found in gluten. Autoantibodies such as anti-transglutaminase characterize the disease as well as intestinal villous atrophy. Several autoimmune comorbidities are known to be associated with CD, such as dysthyroidism or type 1 diabetes. More precisely, the final objective was to estimate the prevalence of the 15 most frequent autoimmune diseases (AID) in a cohort of CD patients followed at HEGP, which is a French national reference center for CD.

Corpus and data processing

Selecting comorbidities from the medical literature. We used the MeSH® (Medical Subject Headings®) co-occurrence file provided by the U.S. National Library of Medicine (NLM) to establish a list of the most frequent comorbidities in CD. We retrieve co-occurrence frequencies of the MeSH main heading Celiac Disease (D002446) with all main headings children of ‘Autoimmune Disease’ (D001327), corresponding to all terms with MeSH Tree Number starting with C20.111.*. We restrained our search to the fifteen most frequent comorbidities.

Selecting the CD corpus. We extracted a CD dataset from our data warehouse. CD cases were identified via an i2b2 query based on the combination of the three following criteria: having had an hospitalization stay with ICD-10 code for CD (K90) in billing claims; one or more stay, or consultation in the gastroenterology department; and at least one text document (discharge or letter) containing the term ‘celiac disease’ or its synonyms. The final corpus was made of all available medical documents for the patient identified by the previous query.

Identifying AID comorbidities in structured data. We leveraged two sources to identify comorbidities in patient: ICD codes from billing system, drugs prescribed and dispensed at the hospital (limited to insulin and levothyroxin). We used the UMLS to map the MeSH terms to the controlled vocabularies used in the CDW, namely ICD10 for diagnoses and ATC for drugs. We used an R script interacting with the voting module from FASTVISU to automatically record identified comorbidities with the corresponding patient.

Preprocessing free text for AID comorbidities identification with FASTVISU. We used a regular expression module for entity recognition. We elaborated a list of regular expressions (regex) for each auto-immune disease. The voting bot using the regex text processing module filtered out documents without any regex match. Expressions and keywords associated with the diseases targeted by the regex were broad to avoid any decrease in sensitivity. We expect the specificity to be near 100% because the set of documents is manually reviewed by experts (establishing a gold-standard).

Review using FASTVISU web interface.

Two trained physicians, with a background in epidemiology and medical informatics, independently reviewed the corpus of documents using the FASTVISU web client. FASTVISU presents all documents from a patient in a chronological order in the left-hand side panel of the screen (see Figure 2). Keywords and expression detected by the regex modules are highlighted. These highlighted concepts are also summarized on the right-hand side panel of the application (and provided with a clickable link to the corresponding occurrence in the text). Each occurrence (e.g. insulin) is associated with a broader category (in this study an auto-immune disease, e.g. diabetes).

After a careful review of set of documents, the reviewer can vote for the presence or the absence of phenotypes and diagnosis. At the end of this step, each patient is characterized by a vote for the presence (or absence) of each of the fifteen diseases, from each of the reviewer (and from the automatic bot voting process).

The next step is the obtention of a consensus between reviewers. This is done in FASTVISU by visualizing only the differences between the two reviewers’ votes. Reviewers will analyze and discuss their choice and agree upon a final decision.
For analysis, the API can export votes. The R statistical software can retrieve poll results directly from the API and can compute prevalence, compare sources of vote (types of codes, human text review).

**Results**

Our cohort contained 741 patients, constituting a corpus of 6340 text documents. The collection is mainly made of consultation summaries and letters (44.9%), and discharge summaries (15.3%). Median number of documents per patient was 5, IQR [3; 10] with a maximum of 146.

The bot filtered the initial corpus, using the regex module, into a corpus containing 847 documents from 276 patients’ records.

Manual review using the web client interface took two hours and two and half hours respectively for each physician. The mutual agreement on the identification of the presence of the 15 auto-immune diseases for the 276 patients was excellent (with a Cohen’s kappa of 0.89).

**Discussion**

*Elements contributing to decrease the review duration.* The human expert time required to manually review was very low thanks to two mechanisms: the automatic selection and pre-identification of relevant entities through regular expressions. First the voting bot filtered records without mention of any of the diseases or concepts related to the disease. Manual review on the remaining documents is fast because the interface presents whole the records on one single page, with links to highlighted occurrences of medical terms of interest. The UI is also conceived to integrate all functionality on a single page to allow fast transition from a patient to the next.

The time spent is neither comparable to retrospective review of paper records, nor to review from the interface of the EHR software, for which the physician would have to find patients one by one, opening every document and voting on a separate application. The best gold standard method, blinded double reviewing \(^2\), is accessible with very reasonable experts effort involved.

Compared to fully automated NLP solutions, human time spent is higher, but human interpretation remains the gold standard, thus confidence in the results is higher with FASTVISU.

Inter-annotator agreement between the two physicians was high. Differences were mostly pointed associated with ambiguous text.

*Technical significance.* The architecture of FASTVISU is flexible. The core REST API can be easily connected to any CDW. In the current architecture, the API queries a relational database (using the i2b2 star-schema). The API could be easily modified to query other types of storage system, including NoSQL databases. In the current version, we used a regex approach know to have a good sensitivity with a potentially low specificity. Thanks to a flexible architecture, adding new “sources” of concept could be easily done (i.e. using external NLP processes to identify concepts in the documents). Our goal is to provide several ways to feed the system (for different purpose).

*Clinical significance.* FASTVISU has been developed at HEGP only recently, but is already used in several studies at the Department of Medical Informatics. It can cover many different cases for a wide range of applications (including patient selection, visit selection, phenotype annotation at patient or visit level).

*Limitation.* The system is at an early stage of development and is not completely integrated in the workflow. We plan to store the results of patient selections in the CDW for later reuse. The system today does not allow voting for single concept occurrences, therefor does not yet allow validating an automatic annotation. Voting on a set of mixed typed entities (text, lab result, care procedure, drug prescription) is not implemented yet, and validating a billing system is not yet feasible. We plan to add these new functionalities in the next version of the software.

*Perspective.* Today, results collected using FASTVISU are not recorded in our CDW. We plan to “close the loop” and to export the results of extractions or patient selection in i2b2. This would allow leveraging concept manually validated in for other usage. We also plan to provide the user with information not only from the text reports but also from other sources, such as structured data, directly in the interface. The combination of complementary sources of information would give a broader and more complete perspective to the reviewers. Our short term perspective is to connect community-issued NLP tools to FASTVISU (as text processing module in FASTVISU, see Figure 1). FASTVISU modular architecture should enable the creation of such connectors easily. Additional automated modules (bots) could
augment the scope of the software. For example, FASTVISU could be used as a workbench to evaluate or to use machine learning algorithms on a set of selected documents.

Conclusion

Cohort selection, phenotypic annotations, validation of concept detected through NLP methods are common tasks in clinical research. FASTVISU allow to easily review set of patient documents, and to vote at different levels of granularity (patient selection in a cohort, encounter selection, phenotype presence or absence). FASTVISU was built upon a REST web-service API, which allows connecting a web-client with several components: a web client, an entity recognition module and a voting module. In this study, we showed that FASTVISU can be used to efficiently detect the presence of auto-immune diseases from a large cohort of patients.

Acknowledgment

This work was supported in part by the SIRIC CARPEM (Cancer Research and Personalized Medicine) program.

References


Supporting Clinical Cognition: A Human-Centered Approach to a Novel ICU Information Visualization Dashboard

Anthony Faiola, PhD MFA1, Preethi Srinivas, PhDc MS1, Jon Duke, MD MS2
1Indiana University School of Informatics and Computing (IUPUI), Indianapolis, IN; 2The Regenstrief Institute, Indianapolis, IN

Abstract
Advances in intensive care unit bedside displays/interfaces and electronic medical record (EMR) technology have not adequately addressed the topic of visual clarity of patient data/information to further reduce cognitive load during clinical decision-making. We responded to these challenges with a human-centered approach to designing and testing a decision-support tool: MIVA 2.0 (Medical Information Visualization Assistant, v.2). Envisioned as an EMR visualization dashboard to support rapid analysis of real-time clinical data-trends, our primary goal originated from a clinical requirement to reduce cognitive overload. In the study, a convenience sample of 12 participants were recruited, in which quantitative and qualitative measures were used to compare MIVA 2.0 with ICU paper medical-charts, using time-on-task, post-test questionnaires, and interviews. Findings demonstrated a significant difference in speed and accuracy with the use of MIVA 2.0. Qualitative outcomes concurred, with participants acknowledging the potential impact of MIVA 2.0 for reducing cognitive load and enabling more accurate and quicker decision-making.

Introduction
Several decades of emerging bedside monitoring devices (BMD) and electronic medical record (EMR) systems have provided intensive care unit (ICU) clinicians with a range of tools that display and intelligently filter data in ways that enhance patient diagnosis. As part of the greater decision-making process, clinical decision-support (CDS) systems have also provided diagnostic tools to identify and analyze patient data through the use of algorithmic rules within a knowledge base. In spite of advances in display technology, the perceptual clarity of visually represented clinical data on LED displays (associated context-sensitive information) continues to be of low quality. (Figure 1.) These constraints have added to ongoing cognitive overload for ICU clinicians, thereby increasing diagnostic error, particularly errors of omission. Correlated to an annual mortality rate of 12-22%, human factors studies have demonstrated that 80% of “user error” is attributable to cognitive overload. These challenges create the potential for missing critical signs of an unrecognized deadly medical condition.

Figure 1. Four interfaces commonly seen in the ICU that present patient data in an array of visual configurations for various purposes: (A) BMD of patient vital signs, (B and C) EMR data and information displays that provide varying levels of clinical decision-support functionality, and (D) Medication pump display (top) and respiratory pump display (bottom).

The Institute of Medicine (IOM) states that the greatest contributor to cognitive load during data extraction and diagnosis is inadequately designed interfaces of bedside devices. In addition to time constraints, work-place interruptions, and BMD alarm fatigue, cognitive filtering strategies are often applied by clinicians as workarounds that severely constrict their analytical capacity. In each case, an assault on the clinician’s visual recognition processor, working memory, and concentration is ongoing due to cognitive loads that exceed human capacity. As a result, the clinical need to process patient information quickly and easily are unmet at critical times of diagnostic decision-making.

Adding to clinician cognitive load is the complexity of the ICU environment, where patients require constant monitoring (by multidisciplinary team members) and support through continuous bedside-care, frequent intervention, and analysis of non-electronic and device-generated diagnostic testing data. Although ICU patients are the most tested and examined of all hospital patients, important medical conditions and physiological deteriorations are (at times) overlooked. In sum, numeric and textual data analysis by clinicians results in excessive cognitive strain and irregular thinking patterns, all of which impact the quality of care and patient safety. The domains of critical care medicine and health informatics have
enormous potential for leveraging and transforming a plethora of patient data/information through the use of human-centered
designed visualization technologies that are smart, mobile, and easy to extract relevant patient-centered knowledge.

With these challenges, this paper is a review of the design and testing of a novel ICU patient information visualization
tool. Intended to support clinician cognitive load and reduce decision-making error, the tool was designed to spatially and
temporally organize contextual patient data. We refer to this tool as MIVA 2.0 (Medical Information Visualization
Assistant, version two. (Figure 2 depicts the MIVA 1.0 and 2.0 interfaces.) Envisioned as a means to rapidly analyze and
interpret trends in EMR data, our goal (in designing the MIVA 2.0 dashboard) was to positively impact patient diagnostic
outcomes and ultimately patient safety by reducing the burden of cognitive strain experienced by many ICU clinicians.

Background

The National Research Council, in its 2009 report, emphasized the need to support cognition, visualization, decision-
making, and workflow optimization in healthcare by means of an effective computing infrastructure.25 Early ICU data
visualization systems were time-oriented that allowed an analysis of electronic patient records, making it easier for
clinicians to quickly assess the overall condition of a patient’s history, while also displaying data trends, significant
information and events, and spot omissions in treatment at several levels of detail and abstraction.33-25

Although critical care decision-support must be facilitated by reliable clinical data,26 information visualization can also
provide valuable assistance to a critical care team for data analysis and patient diagnosis. Notably, effective information
visualization can amplify cognitive processes by providing computer-supported visual representations of patient data. The
purpose of visualization is for rapid information assimilation, pattern recognition, and diagnostic insights derived from
examining large amounts of data. Hence, in addition to existing conventional bedside visual displays, critical care teams
should be supported by appropriate visualization systems in order to reduce user error and ease cognitive load.

Patient data placed in meaningful contexts become relevant medical information that is usable and sharable,27 from which
clinical knowledge can be derived. Through collaborative knowledge sharing, clinical practice allows distributed
intelligence to create a patient-centered community that provides clinical group problem-solving and processes such as
“reflection-in-action.” In other words, ICU information visualization can profoundly impact the predicting of clinical
events, planning the courses of action, and diagnosing patient adverse events. Moreover, the human-centered design and
usability testing of more effective BMD and EMR interfaces continues to be a priority that should support clinical
information processing, patient diagnosis, and long-term patient care.30 As such, this paper reports on a phase-two design
study that compared the usability and effectiveness of a novel visualization application (MIVA 2.0) with a paper charting
system, commonly used to support ICU patient data documentation activities.

System Description

MIVA 2.0 is a novel EMR dashboard technology,31-34 that uses a visualization engine to deliver multivariate biometric
data by transforming it into temporal resolutions. The result is a spatial organization of multiple datasets that allow rapid
analysis and interpretation of trends. Intended as a mobile technology, MIVA 2.0 was not designed to replicate existing
BMD that display patient vital sign data or CDS systems that provide recommendations through rules-based alerts or
predictive models. Rather, MIVA 2.0 was designed to optimize diagnosis speed and accuracy by rapid recognition of
essential changes in physiological data over a designated time frame, e.g., several minutes, hours, days, or weeks. Using
selection menus, ICU clinicians control the necessary data sources and density, time periods, and time resolutions to
narrow down their diagnostic target of a patient’s condition. Figure 2 illustrates the past two iterations: MIVA 1.0 and 2.0.

The MIVA 2.0 system was designed to maximize the clinician’s ability to control and compare what data is visualized
during a specific context-related patient episode or general diagnosis, e.g., during daily rounds. Building MIVA 2.0 as an
interactive prototype (using Flash ActionScript), findings from Study One (MIVA 1.0) informed the redesign of the
visualization system and interface, as illustrated in Figure 2B and Figure 3.

Method

Participants: A convenience sample of 12 clinicians (6 physicians and 6 nurses) of mixed gender from the medical
population of the Indiana University, School of Medicine and School of Nursing were recruited. A bulk email invitation
to potential participants was sent to a group of 15 clinicians known directly/indirectly by the investigators. The 12
volunteers were evenly split into two groups according to gender and profession (physician/nurse), and in the order that
their email arrived confirming their availability for the study. Hence, a control group of six and an experimental group of
six was formed. All participants had various experiences in the ICU and/or Emergency Room. Physicians and nurses
were evenly split between the two groups according to their professional backgrounds.

Study Design: Following the development phase of MIVA 2.0, a mixed-methods evaluation study, comprising of
quantitative and qualitative measures were employed to compare MIVA 2.0 with medical paper records commonly used
for patient charting. Quantitative performance measures were comprised of eight usability time-on-task (speed) test and clinical decision-making (accuracy) task questions and 31 structured usability (Likert scale) questions, and seven qualitative open-ended interview questions. The experimental and control groups completed the usability task questions and interview, while only the experimental group completed the structured usability questions. The experimental group was provided a five-minute priming session to familiarize themselves with the functionality of MIVA 2.0 (Figure 2B), while the control group was provided a five-minute priming session to understand the paper medical chart data (Figure 4).

Participants from both groups were provided a written paragraph that outlined a fictitious clinical scenario to establish the context of the medical event. (See box below.) Both the paper charts and the MIVA 2.0 system reflected the same 12 hours of charting data. In the latter case, MIVA’s patient data was pre-loaded into the system for purposes of the study. MIVA 2.0 data was streamed from a Microsoft Excel data loop that resided on the laptop computer. (Although MIVA 2.0 was designed for use on a laptop and tablet, the usability study used the former because most clinicians noted this as their primary form of reviewing patient information.) Once the participants reviewed the clinical scenario, they were provided the eight multiple-choice questions. (See the Appendix for all usability task questions.) As such, to ascertain the correct answer to the eight questions, the control group analyzed data from the paper medical charts, while the experimental group clicked through the necessary menus/controls of the interactive MIVA prototype to find and analyze data.

Figure 2. (A) MIVA 1.0 static prototype developed for PC Windows (Study: 2008) and (B) MIVA 2.0 interactive prototype developed for tablet and laptop. Core tools/functions include: 1) Vital sign pool, 2) Information-visualization timeline, 3) Data-point scrubber-bar, 4) Intervention/note icon tray, 5) 15 min. data-status, 6) New clinical note generator, 7) Current data-status, 8) Attending clinicians, 9) Time-line indicator, 10) Time resolution control, and 11) Current time /date.

**Clinical Scenario:** A 6-month old infant has undergone repair of an AV Canal. The post-operative course is complicated by pulmonary hypertension, requiring nitric oxide (iNO) to be started on the second postoperative day. On the third postoperative day you are called to the bedside at 14:15 because of an acute deterioration. The bedside nurse states that the patient’s mean arterial pressure (ABP) and mixed venous oxygen saturation (SvO₂) have declined and the mean pulmonary artery pressure (PAP) has increased over the past 15 minutes.

Test Measures and Methods for Data Analysis

*Usability (Time-on-Task) Speed Test:* Participants from the experimental and control groups were tested for execution speed, i.e., time taken to complete each task in minutes and seconds. No time limits were imposed. The Mann-Whitney non-parametric test and independent samples t-test was used to identify any significant differences in task execution speed between the two groups. SPSS v21 was used in all usability statistical analysis.
Usability (Clinical Decision-Making) Accuracy Test: Participants from the experimental and control groups were also tested for the decision-making accuracy of their responses to each of the multiple-choice questions. The response to each question was judged as either correct or incorrect. A Chi-squared test was used to identify the overall significant difference in accuracy between the experimental and control groups.

Post-Test (Context-of-Use) Questionnaire: Following the usability time-on-task and decision-making test, the control group participants were also allowed a five-minute priming session to experiment with MIVA 2.0. (We anticipated that their valued input as clinicians would provide further information (positively or negatively) regarding the potential of the system.) Hence, participants from the experimental and control groups were provided a post-test questionnaire composed of structured and semi-structured questions that addressed the context-of-use and usability of MIVA 2.0. The semi-structured questions focused more on task execution, medium of execution, learning, cognition, articulation, and long-term outcomes of either paper charts (control) or MIVA 2.0 interfaces (experimental). Participants were provided structured Likert questions (Range 1-5) and semi-structured questions, which yielded mean scores, percentages, and short comments regarding MIVA’s interface and interaction design, and overall usability.

Open-Ended (Usability) Interview: Participants from the experimental and control groups were given an interview to comment on the general functionality, usability, utility, and user experience of MIVA 2.0, as well as its potential to support clinical work in the ICU. All 12 participants were interviewed, which were video recorded and transcribed. A content analysis was done of the transcription using ATLAS.ti (v1.0.15). Participants were allowed ample time to discuss the usability, usefulness, and limitations of MIVA 2.0 and its potential use in the context of a real-world ICU environment.

Findings

The usability speed test identified no significant difference in time-on-task between the control group (M=1.30, SD=.78) and the experimental group (M=1.53, SD=.87). However, an independent samples t-test identified that the experimental group (MIVA 2.0) participants performed significantly faster, overall, than the control group in answering two questions: [Q3] t(10)=3.11, p=.011, r=.70; [Q4] t(10)=3.65, p=.004, r=.76. The clinical decision-making accuracy test identified an overall significant difference in accuracy of the eight question test between the experimental (M=.65, SD=.30) and control groups (M=.58, SD=.36): χ²(1,12)=5.04, p=.03, suggesting that the experiment group test score was more accurate than the control group. See Appendix for all eight questions.

The post-test structured questions yielded a mean response that included 67% of both the control and experimental groups reporting that MIVA 2.0 provided consistent interface and interaction sequencing. This included: clarity of wording, meaningful icons, labels positioned appropriately, requiring a minimal learning curve, and had overall positive feelings towards the interface based on their experience. Also, 83% of all participants reported minimal user action required in using MIVA 2.0. Regarding post-test semi-structured responses, participants noted that MIVA 2.0 provided five interface elements or functions that added to its potential as a clinical information dashboard:
1. MIVA 2.0’s information visualization points were easily accessible and without the need to additionally review traditional paper charts,
2. MIVA 2.0’s clinical use was consistent with ICU clinical practice,
3. MIVA 2.0 provided easily understandable icons that represented clinical interventions such as lab work, x-rays, meds, etc., which provide external clues about group work, coordination, and communication,
4. MIVA 2.0 provided easy access to externally distributed knowledge (e.g., clinical notes port in through the electronic medical record system), and
5. MIVA 2.0 was a solution to resolve conflicts about interpreting others’ activity.

NOTE: Seventy-five percent of all participants agreed that current approaches to collecting and presenting ICU critical care data is not sufficient for supporting accurate diagnoses or management of the critically ill. They also made comparative comments about the importance, limitations, learnability, and long-term consequences of paper chart verses MIVA 2.0. See Table 1.

Content analysis of transcribed interviews yielded in-depth participant perceptions of how MIVA 2.0 might successfully or unsuccessfully integrate into the ecological space of the ICU. Five major themes emerged from our qualitative analysis, which we describe in detail below. (Also see the visual representation of these five themes in Figure 6.)

The themes include: (1) MIVA 2.0’s potential for providing current information tools and needs, (2) Current division of labor in the ICU, (3) Implicit and explicit rules, strategies, and checklists currently used in the ICU and MIVA 2.0 complementing them, (4) Changes to the clinical practices and outcomes brought about by MIVA 2.0, and (5) Drawbacks of the changes caused by MIVA 2.0 to the ICU ecological space.

Table 1. Comparison of conventional paper medical charts and MIVA 2.0.

<table>
<thead>
<tr>
<th>TOPIC</th>
<th>PAPER CHARTS</th>
<th>MIVA 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance to Clinical Work</td>
<td>• Absolutely critical in the ICU.</td>
<td>• Can be very helpful in the ICU.</td>
</tr>
<tr>
<td></td>
<td>• Have natural affordances that are difficult to replicate.</td>
<td></td>
</tr>
<tr>
<td>Learnability</td>
<td>• Require an initial investment in learning.</td>
<td>• Is similar to other information displays and thus requires minimal learning.</td>
</tr>
<tr>
<td></td>
<td>• Work routinely in the ICU environment with no learning needed.</td>
<td>• Is much easier than commercial EMR systems.</td>
</tr>
<tr>
<td>Limitations</td>
<td>• Cannot view data efficiently or integrate different pieces of information very well.</td>
<td>• Lack of familiarity.</td>
</tr>
<tr>
<td></td>
<td>• Require more mental effort: flipping pages with different data in different places.</td>
<td>• May only be able to compare five visual data points simultaneously.</td>
</tr>
<tr>
<td></td>
<td>• Are not computable.</td>
<td>• Lack of the undo option.</td>
</tr>
<tr>
<td>Long-term Consequence</td>
<td>• Provide greater chances for errors.</td>
<td>• No emergency exit appears anywhere.</td>
</tr>
<tr>
<td></td>
<td>• Give an inefficient interpretation of data leading to inaccurate conclusions.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• May expedite data interpretation / trend analysis.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Provides user-controlled display with info access.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Allows for faster decision-making.</td>
</tr>
</tbody>
</table>

(1) Current information tools and needs. As a preface to their comments on MIVA 2.0, participants reported that paper and electronic tools were the current standard of use in the ICU. They also noted that strips of paper are used owing to ease of data entry, while health information technology, (e.g., bedside devices and EMR systems) are also used to display and store massive patient datasets in addition to clinician orders. Participants stated that although paper offered a natural affordance and was convenient with respect to recording data, data access and interpretation and sorting through complex data was limited. The visualization component of MIVA 2.0 was reported as easier to interpret at a glance and helped sorting through complex datasets.

Participants also agreed that MIVA 2.0 had the capacity of complementing the current information and cognitive needs of ICU clinicians. Overall, participants showed a positive attitude towards the interface look/feel, interaction sequencing, and the learning curve of the MIVA 2.0. Participants reported becoming familiar with MIVA 2.0 with minimal effort. There were, however, minor differences with respect to the cosmetic interface and interaction sequences between the control and experimental participants. Thus, while the overall look and feel was stated as intuitive, clear and easy to understand by the control participants, the experimental participants remarked that the appearance and behavior of the interface still had several issues to overcome. For instance, control group participant P3 (experimental group) and P4 (control group) stated:
“You need to use color sparingly. I think the use of color is smartly done here. There are not too many different colors. And so the colors are used strategically to facilitate decision-making. The contrast seems appropriate with the colors chosen. The data is easy to manipulate, allowing me to see a different time scale.”

“MIVA should be tested with a scenario with many more data points, e.g., more clinical notes. Most ICU patients, if they are there for more than a week, will have more than a 1000 notes. Can you put 1000 icons up there? ...Maybe you have to group them.”

Participants stated that MIVA 2.0 increased efficiency when compared to current tools with respect to time expenditure and that it contributed to reduction in the cognitive effort, which is needed for data interpretation and decision-making as compared with the current electronic tabular formats and/or paper charts. For example, P6 stated,

“One ends up looking back at each number in a tabular format, similar to paper format. You could do some simple graphing of the data over time, which wasn’t terribly intuitive or usable. One is left with using tabular data. The graphical representation to my eye looks much easier to interpret when comparing variables.”

(2) Division of labor. Participants reported the potential of MIVA 2.0 to support ICU collaborative working patterns (and workflow overall), where multidisciplinary teams of clinicians coordinate during rounds to understand, discuss, and administer patient care. Participants also perceived MIVA 2.0’s potential to support the handing-off process during the change of shifts, commending MIVA 2.0’s ability to off-load manual data-entry of patient information by nurses. For example, P1 stated,

“These data points/graphs across the top, in terms of clinical notes or x-rays taken or medications, allows you to see what data is in relation to each other. For example, if a nurse performs a certain task using certain parameters he/she can get better at understanding the patient data after the fact without having to communicate directly with others about the information. So you have the presentation of all the information from all of those different groups all in one simultaneous place that would allow for enhanced deliver of care.”

(3) Implicit and explicit rules, strategies, and checklists. Participants stated MIVA 2.0 would be useful in following requirement guidelines and checklists (dictated by hospital administration), specifically to ensure that all patient care follows established protocols in administering medications. For example, P10 explained,

“From a physician’s standpoint, it’s the overall care that is important; e.g., if they are on a ventilator we have a checklist to go through to determine if they have had the daily weaning of their sedation. We look at ventilator settings to see if there is an ability to take them off based on a couple of parameters… It (MIVA 2.0) would be helpful for some of our weaning off of the ventilator, i.e., by having information displayed over time you can compare it to previous day.”

(4) Changes in clinical practices and outcomes by MIVA 2.0. Participants reported the shift in clinical practices towards more time being spent on analyzing patient data and making more informed decisions as a result of MIVA 2.0. For example, P6 stated,

“After people get used to MIVA, they will make more educated decisions based on the patient’s status. They would be able to spot trends sooner or more accurately than compared to the tabular data. Then hopefully in the long run, improve patient outcomes… If you are able to take this with you on your rounds on a tablet, this would allow clinicians to be more informed throughout the rounds, thus allowing you to make more accurate or quicker decisions in terms of care management… That would make sense if you were able to see the data, the trends and the relationships easier, i.e., you would be able to address the problems sooner and more efficiently, thereby, hopefully, improving outcomes.”
(5) Drawbacks of changes to the ICU ecological space caused by MIVA 2.0. Although overall participants expressed positive attitude towards the inclusion and usage of MIVA 2.0 in the ICU, they were some who expressed concern over the increased interactivity and access to patient information. For example, P2 contradicted P6:

“...the tradeoff is the extent by which MIVA distracts the team from engaging with one another. I don’t like it when I am working with a distracted team member. I don’t like it when I am working with a resident that is looking up the labs in the middle of discussing the case. So, they may not be engaged to the extent to which they should be if they are using these visualizable tools, i.e., a potential trade-off would be their distraction from engaging with one another.”

Discussion and Conclusion

This paper described the testing of MIVA 2.0, an information visualization dashboard tool for the ICU. A mixed methods approach comprised of performance/usability testing, post-test structure and semi-structured questionnaires, and open-ended interviews were conducted. Findings suggest that MIVA 2.0 has the potential to out-perform the use of paper charts (or other electronic one-way data input devices) in retrieving and analyzing patient data. Participants also noted that MIVA 2.0 was designed with a keen awareness of the broader context for the (real-world) ICU experience; and as such, has the potential to support a rich social matrix of clinical activity. Further, there was concurrence that MIVA 2.0 showed promise for significantly impacting clinical decision-support, as well as improving clinical workflow effectiveness. One participant suggested that adhering to internal workflow practices would also help MIVA 2.0 integrate seamlessly into the current ICU EMR system, provided the visualization paradigms used are carefully balanced between being informative and visually distractive.

All the participants agreed that MIVA 2.0 has the potential to change current ICU clinical practice through the emergence of new analyses and reduced time and effort in placing requests for data. They noted that this might also lead to changes in the function of some of the clinical roles in the ICU environment. The participants, however, suggested that integrating predictive clinical rules and providing a mechanism for the clinician to customize the tool (e.g., setting alerts) would lead to better decision support and patient care. In addition, participants noted scalability and programmed spontaneity as potential challenges that might be faced when MIVA 2.0 is implemented. Additional, however, the authors believe initial tool learning may impact cognitive load.

Limitations of the study relate primarily to this early stage of development of the visualization prototype, a relatively small sample size, and the use of a single clinical scenario. As such, our study should be expanded to include more advanced computer prototypes, a larger sample size, and a broader range of problems, ideally from several clinical ICU settings that require collaboration with intensivists across teams and physical locations. Further limitations of the study relate to lack of consideration for communication and collaboration between the critical care team members in the ICU. For example, during the open-ended interview sessions, participants discussed the need for frequent communication between the team members as a critical component of ICU collaborative processes. For instance, if there is the need for frequent communication between the respiratory therapy team, physical therapy team, nursing team and medical team, MIVA 2.0 could provide specific tool functionality that would facilitate greater interaction and access to workflow information.

Underscoring the importance of these latter findings (related to communication and collaboration), the authors have begun a review of significance of integrating communication-information technology (CIT) into the existing MIVA framework. In further support of a revised model, studies have shown that a major source of workflow error and cognitive strain is related to communication and collaboration breakdown. These findings state that 91% of all medical mishaps are due to communication difficulties and inefficient team collaboration and decision-making. Communication among clinicians, including but not limited to face-to-face interaction, is often interrupted and of poor quality. This has lead to inefficiencies and potential error in the ICU, where rapid and accurate communication is essential for delivering safe patient care. Also, inadequate and inefficient collaboration among nurses and doctors increases the average length of stay of patients, leading to severe inconvenience and greater patient mortality. Direct verbal communication is one of the chief sources of trust building among ICU clinicians, which fortifies work relationships and cognizance of others’ expertise, leading to increased collaboration. In sum, it has been found that ICU clinicians using CIT improve team relationships, staff satisfaction and patient care. Such technologies can improve communication speed by 92%, communication reliability by 92%, coordination by 88%, reduced staff frustration by 75%, and faster and safer patient care. In addition to the above discussion and limitations of MIVA 2.0, MIVA 3.0 is currently being designed to include communication functionality. Since nurses and physicians hold different abilities and experiences of clinical decision-making, good communication and collaboration between multidisciplinary teams is essential. Hence, communication among clinical staff should consist of more than face-to-face, but also incorporate the use of synchronous and asynchronous CIT (e.g., smartphone, email, text and video conference) in order to optimize and enable bi-directional,
rapid, secure, and non-disruptive transmission of content-rich messages, for purposes of expediting and increasing the accuracy and effectiveness of decision-making.45

While the study in this paper focused on usability, user experience, and reducing cognitive load and time spent in analyzing patient data, our assessment only indirectly addressed MIVA’s impact on workflow. For this reason, the author’s current development of MIVA 3.0 includes the goal of optimizing real-time diagnostic solutions during clinical workflow. The primary aim of this work will firstly uncover evidence from three hospital ICUs to more precisely determine how to model workflow, using methods that include ethnographic observation, shadowing and self-reporting interviews. The Experience Sampling method will also be used to collect data on clinician cognitive load and decision-making. From this data, we will construct workflow models that comply with IOM recommendations for ICU patient safety and clinical effectiveness and efficiency.46 ICU workflow models will provide the underlying groundwork for the design of the MIVA 3.0 prototype, followed by a series of comparative (clinical scenario) studies to evaluate the impact of the prototypes on team performance.

Acknowledgement

We thank the Indiana University Purdue University Solution Center for providing funding for MIVA (proof-of-concept) funding for research, Phases 1-4, which supported the design, development, and testing of MIVA 1.0 and 2.0.

References

2. Specific CDS systems currently include computerized alerts/reminders, guidelines, order sets, workflow tools, patient data reports/dashboards, computerized physician order entry, diagnostic support, decision-support systems.
3. Light-emitting diode based video displays.
35. Flash ActionScript is an object-oriented programming language developed by Macromedia Inc., which now resides as part of the Adobe Systems suite of products.
36. ATLAS is a qualitative data analysis program used to do content analysis of interviews and other text-rich content. See: http://atlasti.com/
37. Paper charts, bedside devices, and decision support through remote monitoring.
42. O’Connor C, Friedrich JO, Scales DC, Adhikari N. The use of wireless e-mail to improve healthcare team communication, Journal of the Am Med Info Assoc. 2009;16(5):705-713.
APPENDIX

MIVA In-Lab Scenario / Usability Time-on-Task/Speed and Decision-Making/Accuracy Test Questions
For Experimental Group (MIVA) and Control Group (Paper Chart)

1. What is the current ventilator rate (RR)? a) 16, b) 18, c) 24, d) 22, e) 14
2. What was the mixed venous oxygen saturation at noon yesterday? a) 65%, b) 70%, c) 75%, d) 80%, e) 85%
3. When (day / approximate time) did the pulmonary artery pressure begin to increase? a) Post op day 1, early afternoon, b) Post op day 1, late afternoon, c) Post op day 2, early morning, d) Post op day 2, late evening, e) Post op day 3, Noon
4. When was inhaled nitric oxide commenced? a) Post op day 2 - 0100hrs, b) Post op day 2 - 0800hrs, c) Post op day 2 - 1200hrs, d) Post op day 2 - 1400hrs, e) Post op day 2 - 2000hrs
5. What was the starting dose of inhaled nitric oxide? a) 20ppm, b) 25ppm, c) 15ppm, d) 40ppm, e) 30ppm
6. What happened to the mixed venous oxygen saturation after starting nitric oxide? (Context: Transient is defined as <30minutes; Sustained is defined as >30minutes) a) Sustained decrease, b) Sustained increase, c) No change, d) Transient decrease, e) Transient increase
7. What event precipitated the most recent acute deterioration? a) Withdrawal of iNO b) Ventilator rate weaning, c) Increase of iNO dose, d) Increase of ventilator rate, e) Insufficient data presented
8. How does the current mean PA pressure compare to the mean systemic pressure? a) PA pressure ~1/3 systemic, b) PA pressure is supra-systemic, c) PA and Systemic pressures are equal, d) PA pressure is 2/3 systemic, e) PA pressure is ½ systemic

Post Test - Structured Usability and Context-of-Use - Likert Questionnaire
For Experimental Group (MIVA) [1=SD 2 = D 3 = NA 4 = A 5 = SA]

>> CONSISTENCY <<
1. Was the user feedback consistent?
2. Was the labeling configuration consistent?
3. Was the labeling and/or wording familiar?
4. Was the display consistent with clinical conventions?
5. Were the user actions required consistent?
>> LEARNABILITY <<
6. Did MIVA provide clarity of wording?
7. Was the data grouping reasonable for easy understanding and analysis?
8. Was the grouping of menu options logical?
9. Were the command names meaningful?
10. Were abbreviations, acronyms, and graphic symbols useful and clear?
11. Did MIVA provide meaningful symbols/icons for the verbal labels?
>> MINIMAL ACTION <<
12. Did MIVA require minimal cursor/pointer positioning and action?
13. Did MIVA require minimal steps in sequential menu selection?
14. Did MIVA require minimal user control actions?
15. Did MIVA provide understandable hierarchic menus for sequential selection?
>> PERCEPTUAL LIMITATION <<
16. Were all or most display elements distinctive?
17. Does MIVA provide easily distinguished colors?

Post Test - Open-Ended Interview Questions
For Experimental Group (MIVA) and Control Group (Paper Chart)
1. What was your immediate impression of MIVA?
2. Did you enjoy using MIVA?
3. What did you like the most/least about the MIVA?
4. How would you grade MIVA’s easy of use? A B C D F
5. What was your first impression about the colors, type, interface layout, and interactive tools of MIVA?
6. Did you understand how to use the menus and buttons right away?
7. What did you think were the biggest problems with MIVA?
Automated mutual exclusion rules discovery for structured observational codes in echocardiography reporting

Thomas A. Forsberg, MSc1, Merlijn Sevenster, PhD2, Szymon Bieganski, MD3, Puran Bhagat, MSc2, Melvin Kanasseril, MSc2, Yugang Jia, PhD2, Kirk T. Spencer, MD4
1Philips Research Europe, Suresnes, France;
2Philips Research North America, Briarcliff Manor, NY;
3Philips Healthcare, Andover, MA;
4University of Chicago, Chicago, IL

Abstract

Structured reporting in medicine has been argued to support and enhance machine-assisted processing and communication of pertinent information. Retrospective studies showed that structured echocardiography reports, constructed through point-and-click selection of finding codes (FCs), contain pair-wise contradictory FCs and reliability thereof. In a prospective study, contradictions were detected automatically using an extensive rule set that encodes mutual exclusion patterns between FCs. Rules creation is a labor and knowledge-intensive task that could benefit from automation. We propose a machine-learning approach to discover mutual exclusion rules in a corpus of 101,211 structured echocardiography reports through semantic and statistical analysis. Ground truth is derived from the extensive prospectively evaluated rule set. On the unseen test set, F-measure (0.439) and above-chance level AUC (0.885) show that our approach can potentially support the manual rules creation process. Our methods discovered previously unknown rules per expert review.

Introduction

Structured reporting in medicine has been argued to support the communication process through the use of common report structure and standardized vocabulary, and ideally allows for automated processing and re-purposing of its content to drive downstream clinical, research and operational workflows1,2. A commercial echocardiogram reading and reporting solution (Xcelera, Philips) supports construction of structured reports through selection of site-specific finding codes in an anatomically organized point-and-click interface. A finding code (FC) corresponds to a unique observational or diagnostic statement that has a narrative, one-sentence description, e.g., “Injection of contrast documented no interatrial shunt” and “The prosthetic mitral valve appears to function normally”.

In this paper, we explore if patterns between FCs can be discovered in a large corpus of structured echocardiography reports that represent actionable clinical knowledge. In particular, we shall focus on patterns of the form “FCs A and B are mutually exclusive”. Such patterns may depend on straightforward interpretation of the lexicon, e.g., “Left ventricle is moderately dilated” and “Left ventricle is severely dilated”, or on a more complex understanding of the domain and its pathologies, e.g., “There is severe aortic root dilatation” and “The aortic root is normal in size”.

Rules of the form “FCs A and B are mutually exclusive” can be leveraged to alert the echo reading cardiologist when his/her preliminary report contains both FCs A and B. Automated discovery of such rules can replace or assist a human domain expert in authoring such quality assurance (QA) rules.

Retrospective studies have shown that structured echocardiogram reports contain pair-wise contradictory FCs, which lead to confusing report content3. This, in turn, may negatively affect the referring physician’s perception of the echocardiogram interpretation results, may require additional avoidable communication to resolve the confusion, and, if undetected, may impact patient care.
To detect and resolve contradictions within the report, a plug-in to the echo reporting solution was devised in prior work that compares a set of QA rules against the report’s FCs. If any of the rules is violated, the user is alerted and options are presented to resolve the violation. The plug-in was prospectively evaluated in the clinical workflow at the University of Chicago Hospitals (UCH) in the course of 15 months, showing that on average 2.4 conflicts were detected per echocardiogram and that 83% of reports contained at least one conflict. The QA rule set driving the plug-in was created and maintained manually. This process was labor intensive and required expert resources with deep domain knowledge. Rules created at one site cannot generally be transferred to other sites as FC vocabularies are configurable and therefore locally idiosyncratic. If successful, our methods can be leveraged to support the manual rules creation process by proposing automatically derived mutual exclusion rules.

We will approach mutual exclusion rule discovery as a binary classification problem and address it using machine learning techniques. To separate clinically valid rules from nonsensical rules, we use semantic and statistical methods to characterize potential rules. The semantic methods are motivated by the fact that contradictions between FCs manifest in the combined meanings of their respective descriptions. The statistical methods are motivated by the fact that correct reports do not contain contradictory FCs and the assumption that they are therefore less likely to co-occur than non-contradictory FCs, relative to a priori prevalence.

This paper addresses the following three research questions:

1. Can mutual exclusion rules be discovered using standard machine learning techniques?
2. What is the impact on classification performance of the semantic and statistical characterization methods?
3. Can the proposed methods discover mutual exclusion rules that were not in the ground truth QA rule set?

This work contributes to the research program of exploring and evaluating methods that extract meaningful information and knowledge from corpora of medical information with the purpose of leveraging the extracted patterns to increase the effectiveness and efficiency of healthcare delivery.
Figure 2. Popup that alerts the user that a combination of FCs violates one of the quality assurance rules. This violation can be resolved by selecting one FC and pressing the “Keep Single Selection” button, which results in removal of the other FC. The FCs are presented to the user in their shorthand notation, i.e., internal identifier (“TV600.1”) and brief description (“No TR”).

Methods
Corpus and problem definition
An anonymized corpus of 101,211 structured echocardiogram reports was made available by UCH, in which each report was represented as a list of FCs.

A QA rule set was made available by UCH with the same FC lexicon as the report corpus. This QA rule set had been deployed clinically to prospectively detect inconsistencies in 7,986 reports in the course of 15 consecutive months. The QA rule set contained 580 rules, each was of the form “If {all, at least one} of FCs A₁, ..., Aₙ are contained in the report, then {all, at least one, none} of FCs B₁, ..., Bₙ are contained in the report” and “FCs A₁, ..., Aₙ are mutually exclusive”. Using propositional reasoning patterns, all mutual exclusion rules implied by the QA rules were extracted. For instance, the rule “FCs A, B, C are mutually exclusive” implies that the pairs A-B, A-C, and B-C are mutually exclusive. In this manner, the QA rules generated a multitude of simple rules. In this research, only simple rules of the form “FCs A and B are mutually exclusive” were considered, where A and B were “binary” FCs, i.e., FCs that are either contained in a report or not, as opposed to measurement FCs such as “Left ventricular ejection fraction is ____”. The resulting rule set accounted for part of all violations detected by the complete QA rule set in the prospective study.
Each rule in the initial QA rule set was labelled Mandatory or Suggestion. Violations of Mandatory rules had to be resolved prior to finalization of the report whereas violations of Suggestion rules could be ignored. The rule labelling carried over to the mutual exclusion rule implied by each individual rule.

A rule instance is a rule of the form “FCs A and B are mutually exclusive” for any two distinct FCs A and B. A set of N unique FCs gives rise to N(N − 1)/2 distinct pairs of FCs. A rule instance is Mandatory if it is a Mandatory rule in the QA rule set; it is Suggestion if it is Suggestion in the QA rule set; and, it is unaccepted if it is neither Mandatory nor Suggestion.

We define the QA rule discovery problem as the binary classification problem of distinguishing “positive” rule instances from “negative” rule instances, i.e., the non-positive instances. We shall entertain two definitions of positive instances: 1) the Mandatory rule instances (M mode) and 2) the Mandatory and Suggestion rule instances (MS mode).

A development (“dev”) and a complementing test set were created each comprising roughly half of all labelled rule instances. The dev set was for feature engineering and creating classification models. To ensure that the semantic features were based on only part of the FC descriptions, the dev was created as follows. A random subset of FCs was taken comprising roughly 70% of all FCs, called the dev FC set. The dev set contained all rule instances with FCs A and B, for which both A and B were contained in the dev FC set, and thus contained roughly 50% of all rule instances (0.5 ≈ 0.7 × 0.7).

Feature engineering
A rule instance is characterized as a vector of values, each computed by a feature function (simply referred to as “feature”). The features were created and refined based only on the co-occurrence statistics in the dev set and the narrative descriptions in the dev FC set.

Statistical features
In the context of a corpus of reports, rule instance “FCs A and B are mutually exclusive” gives rise to standard probability metrics based on co-occurrence statistics observed in the report corpus: P(A&B), P(A&not-B), P(not-A&B), P(not-A&not-B), P(A|B) and P(B|A). For instance, P(A&B) is defined as the observed ratio of reports that contain both FC A and B, i.e., the portion of reports in which the rule at hand is violated.

To assess if the observed portion of violations P(A&B) is lower than one would expect if A and B are independent, we include the feature P(A&B)/P(A)P(B), named “O/E”. The product P(A)P(B) is the expected portion of violations, where P(A) and P(B) are the observed portions of reports containing A and B, respectively.

Finally, we include as feature the $\chi^2$ statistic that is computed from the 2 × 2 contingency matrix defined by the co-occurrence counts of A and B. This feature quantifies the level of correlation between A and B.

Semantic features
Identical Modulo Keyword (“IMK”). Boolean feature that returns “true” if the descriptions of two finding codes are identical except for a severity keyword, e.g., “LV moderately dilated” and “LV severely dilated”. Conflicting presence and absence indicators are picked up in a similar fashion, e.g., “There is no apical thrombus” and “There is apical thrombus”.

Word Overlap Ratio (“WOR”). Numerical feature that measures the syntactic similarity between the FCs’ descriptions as the ratio between the number of unique words they have in common and the number of unique words that appear in either of them. Stopwords are eliminated in this count and words are stemmed to account for lexical variation. This feature maps two FCs onto a value in [0, 1], where 0 indicates minimal word overlap (no words in common) and 1 maximal word overlap (identical words).

Anatomical Distance (“AD”). Numerical (integer) feature that measures the distance between the FCs’ anatomical regions as the number of anatomies traversed as blood flows from the upstream to the downstream anatomy in a normal heart. The following anatomies were accounted for: Vena cava → Right atrium → Tricuspid valve → Right
ventricle → Pulmonic valve → Left atrium → Mitral valve → Left ventricle → Aortic valve/Coronary cusp. For instance, the anatomical distance between “Increased E/A ratio for age suggests elevated LA pressure” (Left atrium) and “Cannot exclude aortic valvular vegetation” (Aortic valve) is 3.

Mentions of anatomical regions are recognized in the FC description by means of regular expressions that account for common spelling variations based on the FC descriptions in the dev set and online background resources. If no anatomical region is recognized in either one FC description, AD was −1.

Evaluation

Model evaluation
One Random Forest (RF) classification model was created using the dev set in which the Mandatory rule instances were considered the positive instances (M mode), and another RF model was created in which the Mandatory and Suggestion rule instances were considered the positive instances (MS mode). Each model was then evaluated against the test set in the respective mode. To address the first research question, standard methods of evaluation were used: precision, recall, F-measure and area under receiver-operator curve (AUC). In a similar fashion, two Ripper models were created and evaluated. All models were created and evaluated in Weka (version 3.6.11). The RF classifiers were created using 100 trees, unlimited maximum depth, and \( \log_2(N) + 1 \) features considered at each node, where \( N \) = total number of features. The Ripper classifiers were created using 3 folds (2 for growing and 1 for pruning), 2 optimization runs and minimum weight of 2 for the instances in a rule. To address the second research question, every experiment run was executed another two times with feature vectors constrained to the semantic and statistical features.

False positive review
For each of the four evaluation runs (RF/Ripper and M/MS), the false positive rule instances were collected, i.e., the negative rule instances that were marked as a rule by the classifier. The resulting list of false positives was presented to the attending cardiologist who created the QA rule set (KTS). To address the third research question, the reviewer considered each false positive and assessed if it could be adopted as a rule and, if so, with Mandatory or Suggestion status. The acceptance rate was computed for each classifier as the percentage of accepted rule instances.

Results

Characteristics dev and test set
The corpus consisted of 101,211 structured echocardiogram reports comprising 850 unique FCs, which gave rise to 360,825 (= \( N(N-1)/2 \), where \( N = 850 \)) rule instances. The dev set contained 186,355 (= \( 611 \times 610/2 \)) rule instances based on a randomly selected subset of 611 FCs amounting to 71.9% (= 611/850) of the total number of FCs. The test set contained the remaining 174,470 (= 360,825 − 186,355) rule instances. From the initial QA rule set, 2,485 Mandatory and 658 Suggestion mutual exclusion rules were extracted. Table 3 contains the descriptive statistics of the dev and test set.

Feature properties
In the \( \chi^2 \) analysis of the features in the test set, the features IMK, WOR and O/E are most strongly correlated with the positive instances defined as the Mandatory (M mode) or Mandatory and Suggestion (MS mode) rule instances. The feature \( P(A&B) \) is ranking 11th and last in both modes.

The semantic features rank 1st, 2nd and 5th in decreasing order of \( \chi^2 \) value, in M mode and 1st, 2nd and 4th in MS mode. In both modes, O/E is the highest ranking statistical feature, see Table 2.

Key descriptive statistics of select features on the test set are given in Table 3.

WOR and O/E have an uninterrupted value range in which the Mandatory and Suggestion rule instances outnumber unaccepted rule instances, see Figure 3. For WOR, this interval is [0.6, 1], which covers 0.2% (= 362/174,470) of all rule instances, 16.7% (= 193/1,159) of Mandatory rules instance and 0.0% (= 0/325) of Suggestion rule instances. For O/E, this interval is [0.002, 0.046), which covers 0.2% (= 272/174,470) of all rule instances, 11.5% (= 133/1,159) of Mandatory rule instances and 2.1% (= 7/325) of Suggestion rule instances.
Classification results
The two classifiers were evaluated in both modes on the test set, which was disjoint from the dev set. Performance metrics are given in Table 4. Optimal classification performance is achieved by RF using both semantic and statistical features (F-measure 0.439 and AUC 0.885).

Combining the semantic and statistical features generally results in superior performance compared to characterization of rule instances through either method in isolation. When only semantic features were used, the vast majority of rule instances classified as positive had positive IMK value, i.e., the descriptions of the FCs in the rule were identical modulo a severity indicator (ranging from 87.6% [= 176/201] for RF to 100% [= 176/176] for Ripper both in MS mode). This resulted in high precision values (0.834 to 1, see Table 4) but low recall (0.109 to 0.158, see Table 4).

RF outperforms Ripper on most F-measure scores and all AUC scores. In fact, Ripper’s AUC scores indicate no (0.569 and 0.571, on statistical features, see Table 4) to poor (0.625 and 0.639, on all features) predictive performance. RF’s AUC scores indicate fair (0.721 and 0.732, on statistical features) to good (0.869 and 0.885, on all features) predictive performance.

In all four classifier–mode combinations, precision trumps recall. Highest precision was 0.689 and highest recall was 0.322, both for RF in mode M with all features.

False positive review
Due to a technical insufficiency the false positives were only partially reviewed by the attending cardiologist. The acceptance rates of the reviewed false positives is presented in Table 5 for both classifiers. For RF, 33.0% and 37.6% (= 26.3 + 11.3) of false positives were accepted as clinically valid rules in mode M and MS, respectively. For Ripper, the acceptance rates are slightly higher: 34.9% and 42.9% (= 33.0 + 9.9) in the respective modes.

![Distribution of WOR over rule instances](image1)

![Distribution of O/E over rule instances](image2)

Figure 3. Histogram showing distribution of Mandatory, Suggestion and unaccepted rule instances over value ranges for the WOR (left) and O/E (right) features. In each diagram, the left-most bin has been clipped to enhance interpretability. In the WOR diagram, the Mandatory and Suggestion rule instances outnumber the unacceptable rule instances in the interval [0.6, 1]. In the O/E diagram, the Mandatory and Suggestion rule instances outnumber the unacceptable rule instances in the interval [0.002, 0.046].
Table 1. Descriptive statistics of the dev and test sets.

<table>
<thead>
<tr>
<th>No. of unique FCs</th>
<th>Dev</th>
<th>Test</th>
<th>Dev &amp; Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>611</td>
<td>850</td>
<td>850</td>
</tr>
<tr>
<td>No. of rule instances</td>
<td>186,355</td>
<td>174,470</td>
<td>360,825</td>
</tr>
<tr>
<td>No. of Mandatory rule instances</td>
<td>1,326</td>
<td>1,159</td>
<td>2,485</td>
</tr>
<tr>
<td>No. of Suggestion rule instances</td>
<td>333</td>
<td>325</td>
<td>658</td>
</tr>
<tr>
<td>No. of unaccepted rule instances</td>
<td>184,696</td>
<td>357,682</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Features ranked by decreasing $\chi^2$ statistic in the test set when Mandatory rule instances are considered positive (M mode) and Mandatory and Suggestion rule instances (MS mode), respectively. The "$\chi^2$" in the list of features refers to the feature.

<table>
<thead>
<tr>
<th>IMK</th>
<th>Mandatory</th>
<th>Suggestion</th>
<th>Unaccepted</th>
</tr>
</thead>
<tbody>
<tr>
<td>% true (i.e., identical modulo keyword)</td>
<td>14.9% (173/1,159)</td>
<td>0.0% (0/325)</td>
<td>0.0% (3/172,986)</td>
</tr>
<tr>
<td>WOR</td>
<td>0.2 (±0.2)</td>
<td>0.1 (±0.1)</td>
<td>0.0 (±0.1)</td>
</tr>
<tr>
<td>O/E</td>
<td>2.197 (±50.013)</td>
<td>0.633 (±2.801)</td>
<td>4.816 (±165.586)</td>
</tr>
<tr>
<td>AD</td>
<td>0.052 (±0.234)</td>
<td>0.276 (±0.707)</td>
<td>0.667 (±1.414)</td>
</tr>
<tr>
<td>P(A&amp;B)</td>
<td>73.7% (854/1,159)</td>
<td>62.8% (204/325)</td>
<td>63.3% (109,594/172,986)</td>
</tr>
<tr>
<td>P(A</td>
<td>B)</td>
<td>2.3e-5 (±1.2e-4)</td>
<td>2.1e-4 (±1.1e-3)</td>
</tr>
</tbody>
</table>

Table 3. Key descriptive statistics of select features on the test set

<table>
<thead>
<tr>
<th>Random Forest</th>
<th>Ripper</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>M mode</td>
<td>0.881</td>
</tr>
<tr>
<td>MS mode</td>
<td>0.834</td>
</tr>
<tr>
<td>M mode</td>
<td>0.472</td>
</tr>
<tr>
<td>MS mode</td>
<td>0.473</td>
</tr>
<tr>
<td>M mode</td>
<td>0.689</td>
</tr>
<tr>
<td>MS mode</td>
<td>0.654</td>
</tr>
</tbody>
</table>

Table 4. Performance metrics of the classifiers trained on the instances in the dev set and evaluated on the instances in the test set. The upper part of the tables gives the performance metrics if only statistical features are provided; the lower part if all features are provided. Results are differentiated by the two modes. P = Precision; R = Recall; F = F-measure; AUC = Area under receiver-operator curve.
### Table 5. Percentages of false positives accepted as Mandatory and Suggestion rules in manual review, or not accepted. The acceptance rates are differentiated by classifier (RF or Ripper) and used definition of positive instance (M and MS mode). Due to a technical insufficiency, not all false positives were reviewed. The percentage of reviewed false positives are given in the column “Reviewed”.

<table>
<thead>
<tr>
<th>Classifier–mode</th>
<th>Reviewed</th>
<th>Accepted</th>
<th>Not accepted</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF–M mode</td>
<td>61.3% (103/168)</td>
<td>33.0% (34/103)</td>
<td>6.8% (7/103)</td>
</tr>
<tr>
<td>RF–MS mode</td>
<td>61.0% (133/218)</td>
<td>26.3% (35/133)</td>
<td>11.3% (15/133)</td>
</tr>
<tr>
<td>Ripper–M mode</td>
<td>57.3% (86/150)</td>
<td>34.9% (30/86)</td>
<td>10.5% (9/86)</td>
</tr>
<tr>
<td>Ripper–MS mode</td>
<td>56.5% (91/161)</td>
<td>33% (30/91)</td>
<td>9.9% (9/91)</td>
</tr>
</tbody>
</table>

**Discussion**

**Address research question**

Based on our results, we answer this paper’s research questions as follows:

1. Mutual exclusion rules can be discovered with standard machine learning techniques: An RF classifier retrieved 32.2% (recall) of Mandatory rules on unseen rule instances maintaining precision of 0.689 (Table 4).
2. Combining the semantic and statistical features results in superior performance (F-measure 0.368 to 0.439) compared to characterization of rule instances through either method in isolation (semantic: F-measure 0.197 to 0.217; statistical: F-measure 0.205 to 0.232).
3. Our methods can discover new clinically valid mutual exclusion rules that were not in an extensive manually constructed QA rule set: a domain expert accepted between 33.0% and 42.9% (Table 5) of mutual exclusion rules that were identified as clinically valid rules by the classifiers and that were not in the QA rule set.

Classifier performance for separating Mandatory and Suggestion rules from unaccepted rules (mode MS) is more unreliable than separating Mandatory rules from Suggestion and unaccepted rules (mode M). We hypothesize that this is because Suggestion rules are essentially a different concept than Mandatory rules, as they can co-occur. This seems also to be reflected by the descriptive statistics in Table 3: the average Suggestion values typically sit in between the average values of the Mandatory and unaccepted rules. Mixing Mandatory and Suggestion rules creates a less coherent notion of positive instance, which is inherently harder to learn.

The statistical approach was motivated by the intuition that the FCs in valid mutual exclusion rules co-occur much less than one would expect based on joint probabilities, modelled by feature O/E. The answer to the second research questions shows that rule discovery is not only a statistical problem, i.e., it cannot be reduced to finding the optimal O/E threshold.

**Error analysis**

**False positives**

Among the reviewed false positive rule instances, we found that roughly one third had low co-occurrence ratios (i.e., low P(A&B) and O/E values) but for other causes than conflicting FCs. We identified two root causes for this statistical profile.

The first root cause is reporting convention that prohibit the use of FC combinations. For instance, “There is no gradient across the prosthetic mitral valve” and “There is no mitral stenosis” should not co-occur as a matter of convention as there is an alternative to the latter FC that specifically states absence of stenosis in the prosthetic mitral valve. In the QA rule set, such reporting conventions were not encoded.

The second root cause is (near-)tautological FC descriptions. Inclusion of (near-)tautological FCs is bad reporting style, but does not create conflicts in the report that could potentially confuse the report consumer; e.g., “The mitral valve is not well seen, but is grossly normal” and “There is no significant mitral valve disease”. In addition to not co-occurring as a matter of style, (near-)tautological FCs typically have a substantial portion of words in common, which manifests itself in increased WOR values and makes them more likely candidates for being misclassified as Mandatory or Suggestion rule instances.
False negatives

We identify three feature profiles that were common among the false negatives, i.e., the Mandatory and/or Suggestion rule instances that were misclassified as unaccepted.

First, rule instances with \( \Pr(A \& B) = 0 \) or, equivalently, \( O/E = 0 \). In case the rule instance is valid, this indicates that there are no observed violations in the report corpus; in case the rule instance is invalid, it indicates that the two FCs did not co-occur for other reasons. By definition, \( \Pr(A \& B) = 0 \) if and only if \( O/E = 0 \). Table 3 shows that the latter case is much more prevalent: 109,594 rule instances with \( \Pr(A \& B) = 0 \) were unaccepted versus 854 Mandatory and 204 Suggestion. Hence, because of the overwhelming evidence, if \( O/E = 0 \), the classifier is inclined to mark it as unaccepted. This error category accounted for roughly 74% of false negatives.

Second, rule instances with \( \Pr(A \& B) > 0 \) and negligible expected probability \( \Pr(A)\Pr(B) \), i.e., \(< 1e-5\). In this case, one mere co-occurrence of A and B vastly exceeds the number of expected co-occurrences, simply because the expected number of co-occurrence based on prior probabilities \( \Pr(A)\Pr(B) \) is negligible. This phenomenon greatly offsets the O/E feature: we found examples with \( O/E = 1686.8 \). This ratio is obviously unreliable as it is based on only one single co-occurrence, which may well have been the outcome of a chance process. This error category accounted for roughly 3% of false negatives.

Third, rule instances with \( \Pr(A \& B) \) and \( AD > 0 \). For the instances with \( AD > 0 \), 99.6% are unaccepted; this can be derived from Table 3. In addition, if \( AD > 0 \), the anatomical location descriptions in the FCs do not match. This implies that they have decreased number of words in common, which results in lower WOR values, which is suggestive of the rule instance being unaccepted, see Table 3. Lower WOR values indicates that the rule is invalid, Table 3. This error category accounted for roughly 19% of false negatives.

Real-world implications

Our results show that machine learning methods can discover mutual exclusion rules in the UCH corpus. All features are institute independent and the engine can therefore be applied to unseen report sets in the English language, even if they were written with a different FC vocabulary. This is advantageous as the proprietary echocardiogram reporting solution (Xcelera, Philips) allows for modification and extension of the FC lexicon.

Roughly 70% of rules discovered by the classifiers are clinically valid (e.g., RF’s precision is 0.689 in mode M, Table 4). Although this rate is encouraging, it may be too low to merit uncontrolled auto-population of the facilitated reporting plugin. Various strategies can be adopted to filter out the clinically valid rules. First, the discovered rules can be reviewed manually before adoption. In this scenario, the rules can be presented by decreasing certainty value so that the reviewer or review team can decide to stop validating rules if the clinically valid ones get sparse. Second, clinically invalid rules can be flagged during interpretation and turned off immediately or reviewed periodically.

Our performance analysis shows that roughly 30% of clinically valid rules are retrieved by the classifiers (e.g., RF’s recall on Mandatory is 0.322, Table 4). This rate is rather low, but further research is required to assess its impact on the clinical utility of the rule set. As we saw in the error analysis, the majority of false negative rule instances are never violated in the report corpus, which spans more than 10 years of echocardiogram service, and one can thus reasonably argue that their clinical utility is limited.

Limitations

Due to a technical insufficiency, the classifiers had to be re-trained after manual review of their false positives. There was only a partial overlap between the reviewed rules instances and the final classifiers’ false positives.

Success of our approach is ultimately determined by the number of violations detected by automatically discovered rules, not by the mere number of discovered rules, which is driving the F-measure metric adopted in this paper.

The reporting corpus and the QA rule set originate from one single institution. We simulated unseen data by splitting the rule instances in a dev and test set, but further research is required to assess the extent to which our results generalize to unseen data from other institutions.
The ground truth was based on mutual exclusion rules derived from the QA rule set. The review of false positive rules showed that this rule set is incomplete: there are clinically valid rules that were not marked as such in the ground truth. Incompleteness of the ground truth has potentially impacted our results and likely in an adverse manner.

Future research

Potential steps for future research include classifier refinement and evaluation on unseen report sets. We intend to construct a hybrid classifier composed of a manually created top-level decision tree that has Support Vector Machines (SVMs) optimized for classification in each subspace defined by the tree’s leaves. The decision tree component encodes the high-level reasoning process of the classifiers, whereas the SVMs optimize the decision boundaries for the numerical features in the subspaces defined by O/E and WOR. Evaluation of the classifiers on report data from different institutes is key to assessing generalizability of our results. Furthermore, future implementations can be configured to discover positively correlated FC pairs. Such patterns can be leveraged to auto-suggest FCs based on FCs previously entered in the report, which could save time and reduce mouse clicks.

Conclusions

Based on our results, we conclude that clinical knowledge can be derived from a large corpus of structured echocardiogram reports based on co-occurrence statistics and semantic analysis using machine learning methods. Automatically discovered clinical knowledge items can be leveraged as QA rules that monitor quality of echocardiogram reports either retrospectively or as a part of the reporting workflow. Our work exemplifies how structured reporting content can be utilized for data mining and diagnostic imaging workflow enhancement.

References

Patient Engagement in Cancer Survivorship Care through mHealth: A Consumer-centered Review of Existing Mobile Applications

Yimin Geng, MS, MSLIS, Sahiti Myneni, PhD, MSE
School of Biomedical Informatics
The University of Texas Health Science Center at Houston, Houston, Texas, U.S.A

Abstract

With improvements in early detection and treatment, the number of cancer survivors has been on the rise. Studies suggest that cancer survivors do not often receive proper follow-up care despite existing guidelines. Patient engagement is key to healthy survivorship, and mHealth provides a viable platform to empower survivors with just-in-time personalized support. However, our understanding of existing mHealth solutions in cancer survivorship is limited. In this paper, we use Patient Engagement Framework to investigate existing apps to bridge this knowledge gap. App features are mapped to the framework components to determine the level of engagement facilitated. Ability to record treatment summaries has been found in five out of seven apps examined. While collaborative care and social engagement are found minimally, the majority of features (95%) are limited to information and way finding, e-tools, and interactive forms. Limitations of the existing apps and possible improvements to the framework are discussed.

Introduction and Background

“Cancer survivors” refer to those people who have been diagnosed with cancer and the people in their lives who are affected by their diagnosis, including family members, friends, and caregivers [1]. The number of cancer survivors continues to increase due to the aging and growth of the population and improvements in early detection and treatment. As of January 2014, it is estimated that there are 14.5 million cancer survivors in the United States. By January 1, 2024, that number will increase to nearly 19 million [2]. Unfortunately, the current U.S. health care system is failing to deliver the comprehensive and coordinated follow-up care cancer survivors need. As per the Institute of Medicine report in 2006, too many survivors are lost in transition once they complete the cancer treatment [3]. This population is at risk not only for cancer recurrence, but also for receiving inadequate risk-based and routine preventative health care. In response to this, the Institute of Medicine report in 2006 recommended that every cancer patient should receive an individualized survivorship care plan that includes guidelines for monitoring and maintaining their health to help improve the quality of care of survivors as they move beyond their cancer treatment [4]. The basic components of survivorship care plans (SCPs) include a personalized comprehensive care summary and follow-up plan. SCPs are intended to be a patient-centered communication tool to improve the quality of follow-up care for cancer survivors. Although many SCPs have been developed, the incorporation of these plans into processes of care coordination and the ability of SCPs in empowering patients to self-manage their care has been limited [5, 6].

Mobile health technologies play an important role in helping cancer patients to become active participants in their care. Given the ubiquity of mobile devices and the omnipresent wireless connectivity, mHealth solutions have the capability to provide just-in-time support that is both adaptive and targeted to user needs [7]. mHealth solutions can be used to generate and share patient-centered care planning, to manage late effects with cancer and its treatments, to promote lifestyle and behavioral changes, and to assist survivors with communication with health care providers. Several research projects have examined the utility of mobile tools in supporting breast cancer patients. Healthweaver Mobile project, [8] and My Journey Compass [9] afforded cancer patients the ability to engage with their health, their health providers and their support networks. Although mobile apps have the potential for helping cancer survivors to manage their care, our understanding of the current state of mHealth in cancer survivorship is limited in terms of their ability to empower cancer survivors to self-monitor their care plans and engage in healthy living.

The objectives of our study are threefold: (1) to understand the role of mHealth in the facilitation of patient engagement in cancer survivorship care, (2) to identify the strengths and shortcomings of the current mobile apps geared towards empowering cancer survivors in self-management of their care and overall well-being, (3) to examine the applicability of available engagement framework in the assessment of survivorship mobile apps. In this paper, we specifically employ the Patient Engagement Framework (PEF) to describe the engagement features facilitated by the existing mHealth apps for cancer survivors. In the next sections of the paper, we discuss in detail the various components of PEF, our evaluation methodology, and subsequent results and discussions.
Theoretical Rationale

The Patient Engagement Framework (PEF) is a model created to guide healthcare organizations in developing and strengthening their patient engagement strategies through the use of eHealth tools and resources. The framework was first introduced in November 2012 by the National eHealth Collaborative [10] after a yearlong effort by over 150 top experts in healthcare, technology and human behaviors. PEF can be used as a guidance to evaluate the functionality of cancer survivor survivorship apps that potentially facilitate the adoption of SCPs, patient engagement, self-management of cancer, healthy behavioral change, goal setting and reinforcement, peer support, and assist survivors with communication with health care providers. PEF consists of five phases that begin with providing information and progress toward more active partnering and engagement [11]. The five cumulative phases to engagement that make up the framework include “inform me,” “engage me,” “empower me,” “partner with me,” and “support my e-community.” Each phase is described below along with specific examples relevant to cancer survivors.

- **Inform me phase** is about providing basic information on how to find the facility, providers, and services and with various electronic patient education information sources.
- **Engage me phase** is about engaging patients through health and wellness apps to monitor their diet, weight, and exercise routine and track health data, patient self-entering their health records and data, sending appointment and medication reminders by text messaging.
- **Empower me phase** is about offering secure text messaging with providers and engaging in customized text reminders for daily care, health and symptom assessment.
- **Partner with me phase** is about monitoring patients and tracking patient’s health data remotely.
- **Support my e-Community phase** is about offering online e-community support forums for better information exchange between survivors.

In addition to the phases which determine the level of engagement, the framework has 14 different categories that can help facilitate the intended level of engagement. PEF categories across the five phases include 1) Information and Way-finding, 2) Analytics/Quality, 3) e-Tools, 4) e-Visits, 5) Forms: Printable, 6) Interactive Forms: Online, 7) Integrated Forms: EHR, 8) Patient-Specific Education, 9) Patient Access: Records, 10) Patient-Generated Data, 11) Care Team-Generated Data, 12) Interoperable Records, 13) Collaborative Care and 14) Community Support. Within each of these categories, sample elements were provided as part of the framework to guide the system development process. Please see Figure 1 for detailed information regarding the various components (Phase, Category, and Sample Elements) associated with the PEF. As indicated in Figure 1, patient engagement level is embedded as a cumulative dimension of the framework, i.e. engagement categories in a given phase are specified in addition to the ones available in previous phase (e.g. Empower me engagement= Empower me+ Engage me).

In this paper, we used the basic version of the PEF to develop a rigorous evaluation method that enables the examination of the user engagement features facilitated by existing cancer survivorship applications on mobile platforms.

Methods

Study design

This study presents a descriptive assessment and review of the cancer survivorship applications available from the mobile app stores. Two researchers assessed each of the selected apps using PEF independently to determine phases and features of patient engagement. The features of each app were mapped to these 14 categories to assess the level of patient engagement facilitated by the apps. The app features mapped to a particular PEF category were scored based on the phase of patient engagement: 1 point for a feature with inform me phase, 2 points for a feature with engage me phase, 3 points for a feature with empower me phase, 4 points for a feature with partner with me phase and 5 points for a feature with support my e-community phase. The total PEF score for each app was computed by adding up the points assigned to all features embedded in that app.

Selecting Apps for Review

We searched Apple Store, iTunes, Google Play, Nokia Store, BlackBerry App World and Windows Phone for current available mobile applications using the keywords “cancer survivors”, “cancer survivorship”, “survivor care” and “care plan” in November 2014. Initial search returned 15 related cancer survivorship applications. The inclusion criteria defined for this study are 1) target users are cancer survivors who have completed cancer treatment, and 2) content include either follow-up care plan or late effect management or healthy lifestyle or stress management. Of the 15 apps
retrieved using the search terms, seven applications were identified to meet the criteria. We examined the features of these seven apps to categorize their features using PEF as discussed in the study design. After completing mapping of the applications to PEF, two researchers assessed their reliability of agreement using Cohen’s Kappa measure. The coders had a reliability of $\kappa = .76$, indicating strong consistency and reliability in the coding process.

![Evaluation methodology using Patient Engagement Framework for mHealth apps in Cancer Survivorship](image)

**Figure 1.** Evaluation methodology using Patient Engagement Framework for mHealth apps in Cancer Survivorship

### Results and Discussion

Table 1 shows the general descriptions for the apps. All these seven apps are designed for cancer survivors who have completed their cancer treatment. Among the seven apps, five apps have comprehensive treatment summary or follow up care plans. Three apps provide support infrastructure for late effects management, while two apps were designed for childhood cancer survivors. One application provides a networking platform to connect cancer survivors and caregivers to a network of healthcare providers specializing in cancer survivorship. Of the seven apps, two were developed by hospital entities, one by a research university and four by non-profit organizations with or without university collaborations. A detailed description for each of the apps is provided in Table 1.

We observed 42 features in total for all of these seven applications. On average, there are six features for each app. We aligned all the observed features to the 14 categories within PEF. Overall, the 42 app features mapped to seven of the 14 PEF categories described in the previous sections. Majority of the apps (five out of seven) had the ability to record personal health information and treatment summaries. A detailed distribution of the app features across the engagement categories can be seen in Figure 2.
Table 1. General information about mobile apps for cancer survivors evaluated in this study

<table>
<thead>
<tr>
<th>Apps</th>
<th>Developer</th>
<th>General description</th>
</tr>
</thead>
<tbody>
<tr>
<td>App 1</td>
<td>University Hospital</td>
<td>This survivor app supports patients during survivorship phase. This app includes information on recovering and remaining healthy after cancer treatment. It has interactive tools to keep track of questions to ask doctors. It provides individualized end-of-treatment summary and survivorship care plans. Unique feature is only available to testicular cancer survivors.</td>
</tr>
<tr>
<td>App 2</td>
<td>Non-profit organization</td>
<td>This app connects cancer survivors and caregivers to a network of healthcare providers specializing in cancer survivorship. This app can help people in taking charge of their personal and financial health by providing education and information regarding survivorship care plan and healthy lifestyle.</td>
</tr>
<tr>
<td>App 3</td>
<td>University</td>
<td>This app is for survivors who have completed their treatment for breast cancer. The app enables users to record treatment summary and then email right from their phone to any provider that they might see in the future.</td>
</tr>
<tr>
<td>App 4</td>
<td>Collaboration of Non-profit organization and university research group</td>
<td>This app will help cancer survivors create Survivorship Care Plan. It allows cancer survivors to initiate their own SCP and then complete with the help of provider. The survivor and her/his healthcare team can use this app to coordinate the follow-up care and manage ongoing symptoms.</td>
</tr>
<tr>
<td>App 5</td>
<td>Non-profit organization</td>
<td>This app is to help survivors understand and minimize the side effects of cancer and cancer treatment. It offers nutrition guidance and practical tips to help survivors feel better, maintain their strength, and speed their recovery from common cancer side effects.</td>
</tr>
<tr>
<td>App 6</td>
<td>Collaboration of Non-profit organization and university research group</td>
<td>This app is designed for Adolescent and Young Adult (AYA) cancer survivors (ages 15-39). It is an interactive app to assess health habits and general sense of well-being. It offers personalized tips for being more active, eating better and living a healthier life.</td>
</tr>
<tr>
<td>App 7</td>
<td>Hospital</td>
<td>This app offers childhood cancer survivors the resources, tips, and tools necessary to help minimize potential problems, or late effects, caused by childhood cancer and/or its treatments.</td>
</tr>
</tbody>
</table>

Results indicated that the survivorship app features were mapped to only half of the PEF engagement categories while there are no features to map to the other half. The PEF categories that have features mapped to include: Information and Way-finding, e-Tools, Interactive Forms, Patient-Specific Education, Patient-Generated Data, Collaborative Care and Community Support. These categories support several functions such as:

1) Providing basic information on how to find the facility, providers, and services and with various electronic patient education information sources and tools for care plan creation;
2) Facilitating patients to create care plan, self-enter their patient medical history and manage their medical documents;
3) Tracking health data and lifestyle (diet, weight, and exercise routine); sending appointment and medication reminders by text messaging;
4) Sending text reminders for daily care, health and symptom assessment;
5) Offering social media to potentially support better information exchange among survivors; and
6) Providing support environment through online communities, where cancer survivors can engage in peer-to-peer communication.

Such features can enable users to access, and organize health-related information, while promoting patient-provider communication. Through reminder capabilities, the apps allow patients to comply with care follow-up services as recommended in the survivorship care plans. Online social media and forum capabilities allow users to share their progress with peers and family, while getting exposed to a variety of social support mechanisms [12].
The examination of the app features using PEF (see Figure 3) indicated that the current mHealth solutions for cancer survivors had 38% of features at “inform me” level (16 out of 42 features) and 36% at “engage me” (15 out of 42 features), 21% at “empower me” level. There are fewer features at more active phases, 2.5% at “partner with me” level and 2.5% at “support my e-community” (1 out of 42 features). As shown in Table 2, Patient Engagement Score (PES) was computed by aligning each of the features with the engagement phases described in the PEF. The scores ranged from 7 to 17, with majority of the features facilitating lower levels of engagement. Results indicate that App 5 was rated with the highest PES and seconded by App 1. Interestingly, App 5 is the only app that was rated five stars by users who downloaded the apps in both iTunes/App store and Google Play store.
Table 2. Computation of Patient Engagement Score

<table>
<thead>
<tr>
<th>Apps</th>
<th>Inform me</th>
<th>Engage me</th>
<th>Empower me</th>
<th>Partner with me</th>
<th>Support my e-community</th>
<th>Patient Engagement Score (PES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>App 1</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>16</td>
</tr>
<tr>
<td>App 2</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>8</td>
</tr>
<tr>
<td>App 3</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>7</td>
</tr>
<tr>
<td>App 4</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>9</td>
</tr>
<tr>
<td>App 5</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>17</td>
</tr>
<tr>
<td>App 6</td>
<td>🕹️альн</td>
<td>🕹️ alm</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>11</td>
</tr>
<tr>
<td>App 7</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️альн</td>
<td>🕹️ alm</td>
<td>14</td>
</tr>
</tbody>
</table>

In Table 3, we attempted to depict how the apps performed in each of the PEF categories in terms of the engagement level they facilitated. The heat map highlights the sophistication in terms of the apps’ ability to engage patients with a specific information kind. For Information and Way-finding, most app features are at the “inform me” phase, while for e-tools, app features are at “engage me” and “empower me” phase. For Patient-specific education, app features reached “engage me” phase. There are features for patient-generated data at “empower me” phase, while features for collaborative care are at “partner with me” phase. None of the apps have features such as quality and safety reports on providers and healthcare organizations, or patient-specific quality indications and analytics promoting empowerment and engagement. Of the top three apps that received higher PES score, two apps were developed by a hospital entity. Given that the original objective of the PEF is to enable the health institutions engage their patients, the results are intuitive. Overall, results indicated that majority of the apps are still in initial phases of facilitating patient engagement. There is room to improve in several areas of care facilitation and patient empowerment. Two important areas include (a) harnessing social technologies to gather care information and facilitate peer-to-peer connections that offer emotional and informational support when needed, and (b) development and deployment of consumer-centered care coordination tools that facilitate efficient care transition, family involvement, and self-management of care information. Collaborative access to electronic health records, patient-centered analytics demonstrating physician feedback on patients’ self-management of care plan are other key areas that warrant specific focus to attain high levels of patient engagement.
Table 3. Heat map indicating engagement levels for seven apps in each information engagement category

<table>
<thead>
<tr>
<th>Apps</th>
<th>Information and Way-finding</th>
<th>Interactive Forms</th>
<th>Patient-specific education</th>
<th>Patient-generated Data</th>
<th>Collaborative care</th>
<th>Community Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>App 1</td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
</tr>
<tr>
<td>App 2</td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
</tr>
<tr>
<td>App 3</td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
</tr>
<tr>
<td>App 4</td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
</tr>
<tr>
<td>App 5</td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
</tr>
<tr>
<td>App 6</td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
</tr>
<tr>
<td>App 7</td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
<td><img src="null" alt="Engagement Heat Map" /></td>
</tr>
</tbody>
</table>

**Supplements to the Patient Engagement Framework**

Integrating behavior change theories is vital to engage users in self-management of health. For example, Engaging Consumers in Health with Health Care (ECHC) framework incorporates traditional theories of behavior change in a simplified manner by considering individuals, their social and environmental interactions [13]. However, researchers question the validity of traditional theories in digital health era [14]. Existing behavior change theories may be insufficient to inform mobile intervention development, given the novelty and distinctions in the interaction and connectivity paradigm associated with mobile platforms. It is important to understand the nature of health-related behaviors and evolution of user needs in the context of web-based and mobile platforms through development of multi-disciplinary analytics that offer both depth and scale [15-17]. Based on the lessons learned from this study, we suggest the integration of behavior change strategies [18] using technology development frameworks [19,20] that consider the role of human cognition and social-technical factors surrounding users’ health-related behaviors. Such theory-driven, ecologically-modeled mobile health technologies can form the crux of user-inclusive care management models. Subsequently, the resulting mobile ecosystem can presumably engage the “digital consumer” in self-management of health conditions (e.g. cancer survivorship). In addition, existing patient and user engagement frameworks are limited to the operationalization of engagement attributes as technology features. However, to fully analyze and realize the benefits of such technology accessories, the current realm of engagement models should be expanded to provide standardized assessment criteria that facilitate the evaluation of health outcomes. Such an integrated framework results in mHealth solutions that can potentially promote cancer survivor engagement in their own care, while providing an account of the ways in which such increased engagement may result in superior health outcomes.

**Limitations and Future Work**

The study described in the paper is not without limitations. The search strategy employed limited our focus to the official application stores. A more comprehensive search such as searching scientific literature or app clearinghouse websites [21] may retrieve more apps for cancer survivorship care. This study only examined the features of the cancer survivorship care apps in terms of the HIMSS Patient Engagement Framework. Given the complexity of cancer survivorship care, the framework may not be detailed enough to cover all of mobile features for survivorship care. For
example, based on this definition for “Patient engagement”, that is, “patients, families... and health professionals working in active partnership at various levels across the health care system... to improve health and health care...”[22], the family-centered nature of cancer survivorship has not been captured by the framework. Furthermore, the framework used in the study is aimed at health institutions to develop solutions that are patient-engaging. However, the framework is not completely suitable to identify engagement features for technology solutions that solely involve health consumers. In this paper, we did not intend to evaluate the apps for their effectiveness and performance. The patient engagement score determined in the study does not account for the impact on patient health care outcomes. For example, a feature at Support my e-community level may not necessarily have five times the impact on health outcome than a feature at inform me patient engagement level and more features for a single PEF elements may not necessarily indicate superior engagement levels. Further work on the evaluation of the usefulness and effectiveness of the mobile apps in improving health outcomes are needed. In addition, we have limited our review to mobile platforms, and therefore our paper does not provide an exhaustive evaluation of all the survivorship application that may be available only on web platforms. This study is an attempt to use Patient Engagement framework to evaluate the cancer survivorship mobile app features to understand the ways in which technology facilitates user engagement in self-management of care. PEF is created for healthcare organizations to develop and strengthen their patient engagement strategies. However, the cancer survivorship apps evaluated in this study were developed by a variety of organizations to help survivors engage in their own care. In addition, the Patient Engagement framework provides an account of technology-related functionalities alone, however, user engagement can be facilitated in several dimensions through use of novel inter- and intra-personal interactions distributed among the user, the app, and the environment in which the user interacts with the app [23]. Therefore, sole reliance on the Patient Engagement Framework for the design of consumer-facing mobile apps may not result in the elevation of engagement levels in all these aspects. Broader evaluation frameworks that provide us deeper understanding of the utility and implementation of traditional behavior change strategies, contemporary theories of user interactions are needed for the design and evaluation of the mobile applications that potentially promote cancer survivor engagement in their own care.

Conclusions

Mobile technology holds promise in promoting communication between cancer survivors and healthcare providers, patient engagement, and care coordination [26,27]. However, there are few apps that are targeted to cancer survivors when compared to the growing market of mobile health applications. The development and adoption of mobile apps for cancer survivorship care is still in its infancy. Using Patient Engagement Framework, we have discovered that there is a lot of room for improvement in several areas, specifically, integration of data analytics and dashboard features that establish two-way communication among providers and patients, facilitation of peer-to-peer communication through online community infrastructure inbuilt within the app, and provision of mobile access to patients’ electronic health records. Majority of the already operationalized features (95%) are at beginning or intermediate levels of patient engagement (inform me, engagement me and empower me) in terms of documentation of treatment summary and implementation of cancer survivorship care plan. Most active patient engagement features such as facilitating collaborative care and online community integration need to further evolve. Furthermore, we need to develop a new health consumer engagement model that solely aids in the incorporation of user engagement attributes that capture health-related socio-behavioral constructs, and facilitate outcome-based technology modeling and evaluation. Such an integrated approach will afford a design, development, and evaluation process that results in mobile apps that are consumer-centered, user-engaging, and theoretically-aligned.

References


588
Using Big Data to Evaluate the Association between Periodontal Disease and Rheumatoid Arthritis

Michael A. Grasso, MD, PhD, Angela C. Comer, MPH, Dana D. DiRenzo, MD, Yelena Yesha, PhD, Naphtali D. Rishe, PhD

1University of Maryland School of Medicine, Baltimore, Maryland
2University of Maryland Baltimore County, Baltimore, Maryland
3Florida International University and NOA Data Science Institute Inc, Miami, Florida

Abstract
An association between periodontal disease and rheumatoid arthritis is believed to exist. Most investigations into a possible relationship have been case-control studies with relatively low sample sizes. The advent of very large clinical repositories has created new opportunities for data-driven research. We conducted a retrospective cohort study to measure the association between periodontal disease and rheumatoid arthritis in a population of 25 million patients. We demonstrated that subjects with periodontal disease were roughly 1.4 times more likely to have rheumatoid arthritis. These results compare favorably with those of previous studies on smaller cohorts. Additional work is needed to identify the mechanisms behind this association and to determine if aggressive treatment of periodontal disease can alter the course of rheumatoid arthritis.

Introduction
Rheumatoid arthritis is an autoimmune disease characterized by chronic and systemic inflammatory changes. The specific disease mechanism is not fully understood, but it is believed to involve a combination of genetic and environmental factors. It can involve different parts of the body but principally attacks the joints. It is a disabling and painful condition that can lead to substantial loss of function and mobility. Rheumatoid arthritis has a significant impact on society and affects roughly 1% of adults in the developed world. The prevalence of rheumatoid arthritis increases with age, and the disease is roughly three times more common among women. The treatment includes both medical and non-pharmacologic therapies to control inflammation and prevent joint damage.

Periodontal disease is a chronic inflammatory condition of the periodontal tissues. It is one of the most common oral diseases, affecting roughly half of all adults over age 30. The disease is the result of an opportunistic infection by specific microorganisms in the oral cavity. This leads to a destructive inflammatory process, and ultimately to bone and tooth loss, as well as other sequelae. Good oral hygiene and regular professional cleanings are used to prevent and treat periodontal disease. More extensive interventions can be used to treat refractory disease.

The creation of very large clinical repositories has opened new opportunities for data-driven research. One such repository is the national patient database from the Veterans Health Administration (VA), which is available through the VA Informatics and Computing Infrastructure. For each patient, structured information is available, including demographics, diagnostic codes, outpatient visits, hospital admissions, patient orders, vital signs, laboratory testing, inpatient and outpatient pharmacy data, clinical consults, immunizations, mental health screening, associated physicians, and payment information. The repository also includes unstructured and semi-structured information in the form of progress notes, radiology reports, procedure reports, images, and other clinical narratives.

The VA dataset is one of the largest clinical repositories available, providing detailed patient information for approximately 25 million patients who received care at 152 medical centers and more than 800 outpatient clinics across the United States over the past 15 years. The repository includes more than 4 billion progress notes, 2 billion procedure and imaging reports, 1.6 billion medication fills, and 1.5 billion diagnoses. These data will eventually be combined with genomic data from the Million Veteran Program, which when complete, will be one of the largest genomic databases in the world. The VA data set is one of the best, biggest, and most detailed data repositories for studying disease with unprecedented statistical power.

A number of research projects have explored the relationship between periodontal disease and rheumatoid arthritis. The exact nature of this relationship has yet to be defined, and most studies to date have been performed on relatively small patient cohorts. The two conditions might have a non-causal relationship through a common mechanism, due to shared genetic and environmental risk factors. Alternatively, they could have a causal
relationship, with one condition influencing the course of the other. In this paper, we present the results of our inquiries using the VA repository to study the association between periodontal disease and rheumatoid arthritis.

**Methods**

We conducted a retrospective cohort study to measure the association between periodontal disease and rheumatoid arthritis using a very large clinical repository. This work was supported through a grant from The Arthritis Foundation. Our research was also supported with resources and facilities from the Baltimore VA Medical Center, the Veterans Affairs Informatics and Computing Infrastructure, and the Department of Emergency Medicine at the University of Maryland School of Medicine. Regulatory approval for this work was obtained from the University of Maryland School of Medicine, the Baltimore VA Medical Center, and the Veterans Affairs Informatics and Computing Infrastructure.

The VA repository is housed by the Veterans Affairs Informatics and Computing Infrastructure. A secure workspace is provided through a Microsoft Windows-based remote desktop connection within the national VA network. The data are store in a Microsoft SQL Server database. Other applications available in the secure workspace include Excel, MATLAB, Protégé Knowtator, SAS, SPSS, and R.

The VA repository contains data for all patients who received care from 1999 through the present day. As outlined in Figure 1, we limited our search to the years 1999 through 2012. Our dental cohort consisted of all people in the 25-million-patient VA repository who had at least 4 visits in one of the VA dental clinics for any dental issue. We also limited our cohort to people who had enough clinical encounters to have accumulated at least 100 International Classification of Disease (ICD) codes. This focus ensured that enough dental and other clinical information was present to minimize the possibility of bias due to incomplete data. Within this cohort, we identified patients with periodontal disease as those who had at least 4 dental encounters for the condition, indicated as ICD code 523.x.

Also within the dental cohort, we identified people with rheumatoid arthritis based on those who had the appropriate ICD code (714.0, 714.1, 714.2, 714.4, 714.8, or 714.9) and who had at least one positive serologic test (rheumatoid factor or anti-CCP). Our approach to the diagnosis of rheumatoid arthritis was taken from previous work to verify the accuracy of ICD codes. That earlier work concluded that, within the VA population, ICD code 714 is a sensitive screening tool for identifying patients with rheumatoid arthritis. Combined with a positive serology result, it achieved 88% sensitivity and 91% specificity, and had a 93% positive predictive value.

**Figure 1. Identification of the dental cohort.**

![Figure 1](image)
We excluded people with missing or incomplete information regarding date of birth, gender, location, or clinical encounters. We calculated odds ratios using 2x2 tables to measure the association between periodontal disease and rheumatoid arthritis. We also calculated 95% confidence intervals to estimate the precision of each odds ratio.

**Results**

Half of the patients in the dental cohort had periodontal disease (433,674 out of 865,256). A total of 21,442 patients in the dental cohort (2.5%) had rheumatoid arthritis. Patient demographics are presented in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Patient demographics of the dental cohort.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of dental patients</strong></td>
</tr>
<tr>
<td><strong>Number with periodontal disease</strong></td>
</tr>
<tr>
<td><strong>Number with rheumatoid arthritis</strong></td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>25th percentile</td>
</tr>
<tr>
<td>75th percentile</td>
</tr>
<tr>
<td><strong>Male</strong></td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
</tr>
<tr>
<td>Caucasian</td>
</tr>
<tr>
<td>African American</td>
</tr>
<tr>
<td>Hispanic</td>
</tr>
<tr>
<td>Asian</td>
</tr>
<tr>
<td>Other/Unknown</td>
</tr>
</tbody>
</table>

As shown in Table 2, the odds ratio between periodontal disease and rheumatoid arthritis is 1.42 (95% CI, 1.37 – 1.46). This finding holds for men and women, and for those above and below the median age of the cohort. It also holds for Caucasians, African Americans, and Hispanics. It did not hold for Asians, who represented only 0.7% of the cohort.

| Table 2. Odds ratios for periodontal disease (PD) and rheumatoid arthritis (RA). |
|-------------------------------|-------------------------------|
| **Cohort**                    | **N**                        |
| All patients                  | 865,256                       |
| Age < 44 years                | 455,070                       |
| Age >= 44 years               | 410,186                       |
| Male                          | 806,944                       |
| Female                        | 58,312                        |
| Caucasian                     | 509,394                       |
| African American              | 194,546                       |
| Hispanic                      | 49,926                        |
| Asian                         | 5,955                         |
| **RA+PD+**                    | 12,576                        |
| **RA+PD-**                    | 8,866                         |
| **RA-PD+**                    | 421,098                       |
| **RA-PD-**                    | 422,716                       |
| **Odds Ratio**                | 1.42                          |
| **95% CI**                    | 1.37 - 1.46                   |
| Age < 44 years                | 6,844                         |
| Age >= 44 years               | 5,732                         |
| Male                          | 10,862                        |
| Female                        | 1,714                         |
| Caucasian                     | 7,422                         |
| African American              | 2,876                         |
| Hispanic                      | 864                           |
| **RA+PD+**                    | 4,753                         |
| **RA+PD-**                    | 4,113                         |
| **RA-PD+**                    | 229,472                       |
| **RA-PD-**                    | 214,001                       |
| **Odds Ratio**                | 1.34                          |
| **95% CI**                    | 1.29 - 1.39                   |
| Male                          | 28,425                        |
| Female                        | 1,112                         |
| Caucasian                     | 2,100                         |
| African American              | 27,974                        |
| Hispanic                      | 27,512                        |
| **RA+PD+**                    | 4,113                         |
| **RA+PD-**                    | 4,113                         |
| **RA-PD+**                    | 191,626                       |
| **RA-PD-**                    | 208,715                       |
| **Odds Ratio**                | 1.52                          |
| **95% CI**                    | 1.46 - 1.58                   |
| Male                          | 27,974                        |
| Female                        | 27,512                        |
| Caucasian                     | 248,469                       |
| African American              | 248,078                       |
| Hispanic                      | 395,204                       |
| **RA+PD+**                    | 393,124                       |
| **RA+PD-**                    | 395,204                       |
| **RA-PD+**                    | 208,715                       |
| **RA-PD-**                    | 208,715                       |
| **Odds Ratio**                | 1.41                          |
| **95% CI**                    | 1.37 - 1.45                   |
| Male                          | 248,469                       |
| Female                        | 248,469                       |
| Caucasian                     | 191,626                       |
| African American              | 191,626                       |
| Hispanic                      | 393,124                       |
| **RA+PD+**                    | 393,124                       |
| **RA+PD-**                    | 395,204                       |
| **RA-PD+**                    | 208,715                       |
| **RA-PD-**                    | 208,715                       |
| **Odds Ratio**                | 1.41                          |
| **95% CI**                    | 1.37 - 1.45                   |

Discussion

A growing corpus of evidence points to an association between periodontal disease and rheumatoid arthritis. Most previous reports were based on case-control studies and meta-analyses, all of which had relatively low sample sizes. Our approach was the first large-population study to confirm this association, demonstrating that subjects in our cohort with periodontal disease were roughly 1.4 times more likely to have rheumatoid arthritis compared to other dental patients.

Several hypotheses have been proposed to explain this association. One suggests that chronic inflammation from periodontal disease can trigger systemic factors, which can induce rheumatoid arthritis. Another hypothesis suggests that rheumatoid arthritis is triggered through an autoimmune response to periodontal bacterial byproducts.
A third hypothesis is based on the existence of a shared pathway in the development of both periodontal disease and rheumatoid arthritis15.

Our approach has a number of notable limitations. We identified patients with periodontal disease by ICD codes, and not by chart review or by a direct observation of clinical findings. We attempted to mitigate ICD code accuracy11,16 by requiring at least 4 dental clinical visits for periodontal disease. Other studies were based on patient-reported information17, physician interviews, or review of medical records18, which would be very expensive to perform with the very large VA repository. In future work, we plan to refine our patient cohort by applying text analytics for information extraction from electronic health records, as we and other investigators have done on past efforts19,20,21.

Our definition of rheumatoid arthritis, based on previously reported work11, was shown to be very sensitive in the VA population. However, it was not based on the gold standard for the diagnosis of rheumatoid arthritis diagnosis using clinical and radiographic findings. Another limitation was the exclusion of individuals with negative serologic test results, frequently associated with early rheumatoid arthritis or diagnosis at advanced age22. A recent meta-analysis of reports related to the diagnosis of rheumatoid arthritis documented that anti-CCP antibody or rheumatoid factor positivity had only 78% sensitivity (CI, 76%–80%)23,24. If we had relied on that definition in our study, we would have missed a substantial number of patients who actually met our criteria.

Finally, our research was limited to the national VA repository. About 90% of the patients whose records are stored in that database are male and about 60% are Caucasian. In addition, our analysis was limited to U.S. veterans who received both medical and dental care through the Veterans Health Administration.

Additional research is needed to explore the significance of this association and of any temporal significance between the two conditions. These investigations could ask whether the treatment of periodontal disease in people at risk for rheumatoid arthritis (e.g., females, older people, those with specific human leukocyte antigen genes) can prevent rheumatoid arthritis. Research is also needed to study whether the treatment of periodontal disease in people with rheumatoid arthritis can relieve its manifestations (fatigue, joint pain, morning stiffness). Finally, research is needed to identify ways to apply big data analytics to large clinical repositories to elucidate more about this association, and to explore temporal correlations between rheumatoid arthritis and periodontal disease.

**Conclusion**

We conducted a retrospective cohort study to measure the association between periodontal disease and rheumatoid arthritis. This was the first such study to employ a very large clinical repository. We demonstrated that patients with periodontal disease were roughly 1.4 times more likely to have rheumatoid arthritis compared to other dental patients. These results compare favorably with previous studies on smaller cohorts. Additional work is needed to identify the mechanisms behind this association and to determine if aggressive treatment of periodontal disease can alter the course of rheumatoid arthritis.

**References**


19. Dhariwal D, Joshi A, Grasso MA. Text and ontology driven clinical decision support system. AMIA Annu Symp Proc. 2013


Simulation-based Evaluation of the Generalizability Index for Study Traits

Zhe He, PhD¹, Praveen Chandar PhD¹, Patrick Ryan, PhD¹², Chunhua Weng, PhD¹
¹Department of Biomedical Informatics, Columbia University, New York, NY USA; ²Janssen Research and Development, Titusville, NJ USA

Abstract

The Generalizability Index for Study Traits (GIST) has been proposed recently for assessing the population representativeness of a set of related clinical trials using eligibility features (e.g., age or BMI), one each time. However, GIST has not yet been evaluated. To bridge this knowledge gap, this paper reports a simulation-based validation study for GIST. Using the National Health and Nutrition Examination Survey (NHANES) data, we demonstrated the effectiveness of GIST at quantifying the population representativeness of a set of related trials that differ in disease domains, study phases, sponsor types, and study designs, respectively. We also showed that among seven example medical conditions, the GIST of age increases from Phase I trials to Phase III trials in the seven disease domains and is the lowest in asthma trials. We concluded that GIST correlates with simulation-based generalizability results and is a valid metric for quantifying population representativeness of related clinical trials.

Introduction

Randomized controlled trials (RCTs) have been widely regarded as the gold standard in medical research [1]. To ensure the internal validity of a clinical study when testing the efficacy of a treatment, clinical trialists often use restrictive eligibility criteria for participant selection [2]. However, unjustified exclusion criteria may unfairly deprive the opportunity of patients to benefit from the trial, and more importantly, compromise the generalizability of its results to the real-world patient population [3]. Consequently, many FDA-approved medications were later withdrawn from the market due to safety problems that had not been detected in pre-marketing clinical trials but only apparent after exposing the medications to a broader patient population [4, 5].

To assess the population representativeness of clinical trials, researchers may compare the study population of a single trial with a convenience sample of the real-world patient population [6, 7]. Most of the generalizability assessment studies identified in the literature focus on posteriori generalizability and thus can be conducted only after the conclusion and publication of a trial. In contrast, priori generalizability, whose focus is on the eligibility of participants, can be assessed during trial design. Posteriori generalizability is almost always lower than priori generalizability because eligibility criteria always subsume the characteristics of study participants [8]. In addition, our previous study found that many trials, especially those on the same medical condition, often use similar or identical eligibility criteria, indicating that the generalizability issue may not be only at the individual trial level, but also across a whole body of trials for a clinical domain at the research community level [9].

To facilitate a priori generalizability assessment, a method was recently published for systematically assessing the population representativeness of a set of related trials. This method compares the aggregate target populations of all the trials under consideration, which characterize the patients who can be enrolled in these trials according to the inclusion and exclusion criteria, with the “real-world” patient population from electronic health records (EHRs) [10]. The Generalizability Index for Study Traits (GIST) was initially introduced along with this method [10]. This paper reports the initial study evaluating the validity and effectiveness of GIST.

Background

GIST is a mathematical function for quantifying the collective population representativeness of a set of clinical trials with reference to the real-world patient population measured by a single quantitative eligibility feature. There are three parameters in the GIST function: i.e., the real-world patient population (PP), the target population (TP) of a trial set, and an eligibility feature (not shown in the formula for brevity). The GIST metric is conceptually similar to Weisberg et al.’s model of patient selection bias in a trial using a counterfactual framework, which takes into account the proportion of patients with an adverse event incidence and the probability of such patients being selected by a trial [11]. GIST score ranges between 0 and 1, with 1 being most generalizable and 0 being least generalizable. It first discretizes the value range of an eligibility feature into consecutive non-overlapping value intervals and then sums the percentage of studies that recruit patients in each interval multiplied by the percentage of patients in the real-world population observed in that interval across all the intervals. GIST can quantify the representativeness of the target population (TP) for the patient population (PP) with respect to an eligibility feature such as age. The mathematical formula of GIST is:

\[
\text{GIST} = \sum_{i=1}^{n} p_i \times g_i
\]

where \(p_i\) is the proportion of patients in the \(i^{th}\) interval, \(g_i\) is the percentage of studies that recruit patients in that interval, and \(n\) is the number of intervals.
Clinical trial target population (P0): the patients being sought as defined in the clinical trial eligibility criteria.

“Real-world” patient population (P1): the patients to whom the results of clinical trials are intended to be applied. We can only approximate its definition given available data resources about these patients.

Weighted “real-world” patient population (P2): the patients sampled from the “real-world” patient population based on the percentage of trials considering these patients for inclusion.

Clinical trial study population (P3): the study participants who are actually enrolled in a clinical trial. Compared to P1 and P2, the study population P3 maximally (if not perfectly) reflects the target population (P0) because all enrolled study subjects should meet the eligibility criteria that define the target population.

For example, diabetes research may target Type 2 diabetes mellitus (T2DM) patients by defining P0 as “patients with HbA1c above 7.5%”, while the real-world T2DM patients (P1) may be those patients whose HbA1c is above 7.0%, and the clinical trial study population (P3) may be a subset of real-world diabetes patients whose HbA1c is above 8.0%. Therefore, P1 subsumes P0, which further subsumes P3.

In [10], to reveal the population representativeness problem at a research community level, we aggregated multiple related clinical trials and used the distribution of trials over HbA1c values to represent the collective target population, i.e., the percentage of trials considering patients with a certain HbA1c value. In this work, we created P2 by sampling real-world patients based upon the percentage of trials considering them, thereby bringing the real-world patient population (P1) closer to the target population (P0). As such, P2 should be better represented in the collective target population (P0) than P1. Using P0 as the target, increasing generalizability will be observed in P1.

Methods

A conceptual framework for validating GIST

Ideally, the GIST metric can be validated by taking two or more trial sets with known different generalizability with respect to an eligibility feature and assessing if the difference in their GIST scores correlates with the expected generalizability differences. As illustrated in Figure 1(a), given TP1 and TP2 such that the population representativeness of TP1 is known to be better than TP2 with respect to a certain eligibility feature, GIST can be validated by verifying if GIST(TP1, PP) > GIST(TP2, PP). However, it is not feasible to obtain TP1 and TP2 because we have no evidence yet what kinds of trials have better generalizability with respect to a certain eligibility feature. Meanwhile, each trial set is affiliated with three different patient populations, i.e., the real-world patient population, the target population constructed from eligibility criteria descriptions, and the study population that includes all enrolled patients in the trial. Our method for validating the GIST metric is to simulate different patient populations that would result in a known generalizability difference with respect to the same trial set and assess if the difference of their GIST scores correlates with the expected difference. Therefore, we simulated a patient population that has better generalizability than the real-world patient population for the same trial set through weighted sampling of the real-world patient population. The relationships of these populations in our simulation-based validation method are illustrated in Figure 1(b) and their definitions are provided as follows:

\[
GIST(TP, PP) = \frac{\sum_{i=1}^{N} I[l_{\text{low}}, l_{\text{high}}]}{T} \times \frac{\sum_{i=1}^{P} I[l_{\text{low}} < y < l_{\text{high}}]}{P}
\]

where N is the number of distinct value intervals of the quantitative eligibility feature under consideration, T is the number of trials in the trial set included for aggregate analysis, P is the number of patients in the patient population PP, \(w_i\) is the inclusion value interval of the quantitative feature for the \(j^{th}\) study, such that indicator I can be defined as \(j^{th}\) study interval subsumes the \(k^{th}\) interval’s low and high boundary, and \(y_k\) is the observed value of the quantitative feature for the \(k^{th}\) patient such that an indicator I can be defined when \(k^{th}\) patient has a value of the quantitative feature falling within the \(j^{th}\) interval. Note that the GIST metric can also be applied to categorical variables, whereby the value intervals are integers.
P2, and P3, in order. Therefore, if the GIST scores for them follow GIST(P0,P1) < GIST(P0,P2) < GIST(P0,P3), we can conclude that GIST is a valid metric in quantifying the population representativeness of a trial set.

We first validated the GIST metric using this simulation-based method. Then we compared the GIST scores of trial sets that differ in their disease domains, sponsor types, study phases, and study designs. We hypothesized that:

**Hypothesis #1:** Weighted “real-world” patient population can serve as a good reference standard for validating GIST’s suitability for indicating the population representativeness of a set of related clinical trials, one eligibility-feature each time.

**Hypothesis #2:** GIST correlates with the population representativeness of a set of related trials.

To profile the patient populations used for GIST evaluation, we used the population health data from the National Health and Nutrition Examination Survey (NHANES), a continuous cross-sectional health survey conducted by the National Center for Health Statistics of Centers for Disease Control and Prevention (CDC) [12]. NHANES evaluates a stratified multistage probability sample of the non-institutionalized population of the United States. The survey samples are first interviewed at home, followed by a physical and a laboratory test in a mobile examination center. Its rigorous quality control standards ensure national population representativeness and high-quality data collection.

**Figure 2** shows the data collection and analysis pipeline employed in this study. We first extracted patient data from the NHANES database downloaded from the CDC website. We then retrieved the clinical trial summary text from ClinicalTrials.gov. We extracted and aggregated baseline characteristics of enrolled patients in T2DM trials with results. All the data were extracted with R and Python scripts, and subsequently stored in a MySQL database. After processing the data, we first evaluated the GIST metric. Then, we used GIST to compare population representativeness of trial sets of various characteristics. We will explicate each step as follows:

**Step 1: Extracting patient data from NHANES databases**

To ensure the statistical power of the analysis, we identified seven medical conditions in NHANES, each having more than 1,000 samples after combining data in multiple survey cycles. They were Type 2 diabetes mellitus (T2DM), asthma, arthritis, depression, sleep disorders, heart attack, and stroke. In the following, we will describe how we extracted and processed the NHANES data for these seven selected medical conditions.

**T2DM:** We combined the results of the Diabetes questionnaire of five continuous survey cycles between 2003 and 2012 and identified 3,304 diabetics who had their diabetes confirmed by a health professional and one HbA1c (Glycohemoglobin) measurement. As NHANES does not distinguish between two subtypes of diabetes, we employed a method used by Dodd et al. [13] to further identify 3,082 T2DM patients after excluding 222 samples with Type 1 diabetes who were (1) first diagnosed with diabetes before age 30; and (2) taking insulin. The rationale is that as one grows older, his/her lifestyle (e.g., dietary habits) will play a more important role in developing T2DM. Three quantitative eligibility features that are frequently used in T2DM trials, i.e., age (99.0%), HbA1c (53.6%), and Body Mass Index (BMI) (46.6%), were used for GIST evaluation. We combined the laboratory test results on HbA1c and examination data on body measures for five continuous survey cycles between 2003 and 2012. Out of the 3,082 T2DM samples, 2,695 had no missing values for age, HbA1c, and BMI. We used Chi-square test on two categorical variables, i.e., “gender” and “ethnicity”, to test the representativeness of these 2,695 patients with no missing values for all the 3,082 patients. No statistically significant difference was found ($P$-value > 0.05). Therefore, we concluded that these 2,695 patients is a representative sample of all the T2DM patients in NHANES and included them in our further analysis.

**Depression:** NHANES started to conduct interviews on depression in the survey cycle of 2005-2006. We combined the results of the Depression Screener questionnaire of four continuous survey cycles between 2005 and 2012. The
Depression Screener questionnaire uses a 9-item screening instrument that asks questions about the frequency of symptoms of depression over the past 2 weeks. For example, participants were asked how frequently they “have little interest in doing things” or “feel tired or have little energy.” Every question was rated from “0” to “3”, where “0” means “not at all” and “3” means “nearly every day.” Employing a method used by Xiao et al. [14], we identified 1,884 depressive participants who have a combined score of 10 or higher for the nine questions.

**Sleep disorders:** NHANES started to conduct interviews on sleep disorders in the 2005-2006 survey cycle. We combined the results of the Sleep Disorders questionnaire of four continuous survey cycles between 2005 and 2012 and identified 1,816 participants who were told by a doctor or a health professional to have sleep disorders.

**Asthma, arthritis, heart attack, and stroke:** After combing the results of Medical Conditions questionnaire of five continuous survey cycles between 2003 and 2012, we identified 7,009, 7,449, 1,235, 1,119 participants who were told by a doctor or health professional to have asthma, arthritis, heart attack, and stroke, respectively. It is worth noting that NHANES does not provide other laboratory tests to further validate these conditions.

To account for oversampling, non-response, and post-stratification, NHANES assigned each participant a two-year sample weight (WTMEC2YR), which is the number of people in the U.S. national population that each participant can represent. According to the analytical guideline of NHANES [15], we calculated eight-year sample weight WTMEC8YR (1/4 * WTMEC2YR) for depression and sleep disorders patients, because their data in four survey cycles were combined. For the other five conditions, we calculated ten-year sample weight WTMEC10YR (1/5 * WTMEC2YR), because data in five survey cycles were combined for them. After applying the normalized sample weights in the analysis, the patients in NHANES can represent the U.S. non-institutionalized population in the midpoint of the combined survey period.

**Step 2: Retrieving trials data from ClinicalTrials.gov**

To facilitate large-scale systematic analysis of population representativeness of related clinical trials, we have built a computable repository of clinical trials called COMPACT, which stores fine-grained eligibility features and descriptive characteristics of all the trials in ClinicalTrials.gov [16]. COMPACT indexed trials by medical conditions, allowing efficient aggregate analysis of trials on the same condition. Based on COMPACT, we have built a Web-based visual analytic tool of eligibility features in clinical trials called VITTA [17]. VITTA allows its users to select trials of a particular medical condition, refine the selection of trials by various characteristics, and profile the collective target population with a single eligibility feature.

For each medical condition, from COMPACT we retrieved interventional studies that had their start date falling in the survey years of NHANES. Corresponding patient data were obtained. For example, because patient data with sleep disorders were obtained from NHANES between 2005 and 2012, we also retrieved interventional studies on sleep disorders with the start date between January 2005 and December 2012 from the COMPACT database.

**Step 3: Retrieving a convenience sample of enrolled patients in T2DM trials**

As the study population of enrolled patients should well represent the target population of a trial, the collective study population should yield better population representativeness than the “real-world” patient population. To test whether GIST score can reflect this expected difference, we retrieved the results data of T2DM trials between 2003 and 2012 that reported summary data of their enrolled patients in ClinicalTrials.gov. The summary data must report the number of participants, mean and standard deviation (SD) value of at least one of age, HbA1c, and BMI to be included. We aggregated the mean and SD of age, HbA1c, and BMI separately using the following formula (adapted from [18]), where $\bar{T}$ is the number of studies,

$$\text{Weighted}_\text{mean} = \frac{\sum_{i=1}^{T}(\text{mean}_i \times \text{number}_i)}{\sum_{i=1}^{T} \text{number}_i} \quad (2)$$

$$\text{Weighted}_\text{SD} = \sqrt{\frac{\sum_{i=1}^{T}(\text{SD}_i^2 \times (\text{number}_i - 1))}{\sum_{i=1}^{T} \text{number}_i - 1}} \quad (3)$$

Table 1 shows the number of T2DM trials that reported summary data of age, HbA1c, and BMI for their enrolled patients, the total enrollments of these trials, and aggregated mean and SD values of age, HbA1c, and BMI, respectively. Given that only a small number of trials reported summary data of their enrolled patients in ClinicalTrials.gov, these aggregated data represent a convenience sample of enrolled patients in all the T2DM trials.
Table 1. The mean and SD of age, HbA1c, and BMI of the convenience sample of enrolled patients in T2DM trials.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of trials</th>
<th>Number of patients</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>389</td>
<td>198,050</td>
<td>58.3</td>
<td>9.4</td>
</tr>
<tr>
<td>HbA1c</td>
<td>137</td>
<td>62,931</td>
<td>8.2</td>
<td>1.0</td>
</tr>
<tr>
<td>BMI</td>
<td>108</td>
<td>70,678</td>
<td>30.5</td>
<td>5.2</td>
</tr>
</tbody>
</table>

**Step 4: Evaluating GIST using simulation**

From COMPACT, we retrieved 2,731 interventional studies on T2DM with a start date falling between 01/2003 and 12/2012. The number of T2DM trials specifying permissible values for age, HbA1c, and BMI was 2,702 (99.0%), 1,463 (53.6%), and 1,274 (46.6%), respectively. We formed one trial set for each of the three features and included them for generating the distribution of trials (P0) over age, HbA1c, and BMI, respectively. These distributions were used for evaluating the GIST metric. For each feature, we generated three patient samples as follows:

**“Real-world” patient population (P1):** A random sample of 10,000 patients from the T2DM patients in NHANES using normalized NHANES sample weight with replacement. NHANES sample weight is the number of patients in the U.S. national population that one survey participant can represent. Therefore, this sample can represent the “real-world” T2DM patients.

**Weighted “real-world” patient population (P2):** We generated a random sample of 10,000 patients from T2DM patients in NHANES considering both NHANES sample weight and the percentage of trials that consider such patients. Specifically, we separately normalized NHANES sample weight (i.e., WTMEC10YR) and percentage of trials (P0), and then used the average of these two normalized weights to sample the T2DM patients in NHANES. As such, the sample of weighted “real-world” patients generated by oversampling “real-world” patients who are considered by more trials and under-sampling “real-world” patients who are considered by fewer trials would be better represented in the target population of the trials (P0) than the “real-world” patient population (P1). Note that the values in both the normalized sample weights and normalized percentages of trials added up to 100%, while the sum of values in P0 did not.

**Study population of clinical trials (P3):** A random sample of 10,000 enrolled patients generated using Gaussian distribution with the mean and SD of the convenience sample of enrolled patients in T2DM trials (from Step 3).

In the ideal situation, the study population should represent the target population, assuming the enrolled patients match the eligibility criteria perfectly. Therefore, P3 should have the best generalizability of the target population among P1, P2 and P3. The same trial set should have better population representativeness for P2 than P1. For each feature (i.e., age, HbA1c, and BMI), we calculated the GIST scores for three patient samples. Taking sampling variability into account, we ran the experiment for 100 times. Note that P0 for each feature remained the same in all the experiments, whereas P1, P2, and P3 were generated once for each feature in an experiment. If the calculated GIST scores in all the experiments consistently follow $\text{GIST(P0,P1)} < \text{GIST(P0,P2)} < \text{GIST(P0,P3)}$, GIST correlates the expected difference of three patient samples and is therefore a valid metric for assessing the population representativeness of a given patient population in a given trial set.

**Step 5: Comparing population representativeness of trial sets of various characteristics**

To reveal the population representativeness problem at the research community level, we calculated the GIST score of age for each of the seven previously selected conditions: T2DM, depression, asthma, sleep disorders, arthritis, heart attack, and stroke. To compare population representativeness of different types of trials, we further performed stratification analysis on study phases, sponsor types, and study designs across multiple conditions. We used the GIST scores of age for trial sets that differ in these trial characteristics to compare their population representativeness.

**Results**

**Evaluation results of GIST using simulation**

For each of the three eligibility features (i.e., age, HbA1c, and BMI), we generated three patient samples (i.e., P1, P2, and P3, defined in Step 4 of the Methods Section) that have known differences in generalizability for collective target population (P0) of the same set of trials in one experiment and ran the same experiment for 100 times. Even though the difference between P1 and P2 may be minor if most trials accept broad range of values, the GIST metric should still capture the difference. To illustrate the differences among three patient samples, we visualized in Figure...
the distribution of P1, P2, and P3 against the target population of T2DM trials (P0) for age, HbA1c, and BMI in the pilot experiment. The widths of value intervals for age, HbA1c, and BMI are 1, 0.5, and 1, respectively. In each sub-figure, the green dot-and-dashed curve represents the target population of T2DM trials, i.e., the percentage of trials allowing a value interval. The blue solid curve represents the distribution of patients in the sample of “real-world” patients (P1) over consecutive non-overlapping value intervals of a feature. The red dotted solid curve represents the distribution of patients in the weighted sample of “real-world” patients (P2). The light blue solid curve with big dot represents the distribution of patients in the sample of enrolled patients in T2DM trials (P3).

There are two y-axes: the left one is for the three sample patient populations (i.e., P1, P2, and P3) and the right one is for the target population of clinical trials (P0).

Figure 3. Visualization of three patient populations (P1, P2, and P3) and the target population of T2DM trials (P0) with respect to (a) age, (b) HbA1c, and (c) BMI, respectively.
As can be seen in Figure 3(a), both distributions of P1 and P2 peak at age 80, which is considered by only about 40% of trials. The distributions of P1 and P2 are similar in the visualization, but a statistically significant difference between them was observed in two-sample Kolmogorov-Smirnov test (test statistic = 0.028 > 0.014 = critical value, \( P \)-value < 0.05). Figure 3(b) shows the visualization of three patient samples and collective target population regarding HbA1c. The distribution of P3 aligns with P0 better than that of P1 and P2. The peak of the distribution of P2 stands in between that of P1 and P3, confirming that the same set of trials does have better generalizability for P2 than P1. Figure 3(c) shows the visualization for BMI. We can see that all three distributions of P1, P2, and P3 peak at about 30 kg/m\(^2\), where the peak of P2 is higher than that of P1. As BMI value increases from 30 kg/m\(^2\), the curve of P2 gradually drops below that of P1. The reason is that the weight (i.e., target population) we used to generate P2 is set at about 30 kg/m\(^2\). WTMEC10YR or WTMEC8YR) of the survey participants with the corresponding condition listed in the year range. The GIST score of age for trials on the other five conditions ranged between 0.74 and 0.83, showing relatively good population representativeness.

To uncover the population representativeness issue at the research community level, we used GIST to compare the GIST scores of age for seven medical conditions ranked in ascending order (Column 5). The numbers of patients in the U.S. national population (Column 3) are the sum of the normalized sample weights (i.e., WTMEC10YR or WTMEC8YR) of the survey participants with the corresponding condition. The GIST score of age can assess how the target population of a set of trials using age as an eligibility criterion (Column 4) represents the “real-world” patient population (Column 3). We observed that among all the conditions, asthma trials had the worst population representativeness (0.54), while heart attack trials had the best population representativeness (0.89). The GIST of age for trials on the other five conditions ranged between 0.74 and 0.83, showing relatively good population representativeness.

Roumiantseva et al. have found that industry-sponsored studies differ systematically from government-sponsored studies in study type, interventions, and condition studied [19]. We are also interested in exploring the difference of population representativeness for trial sets of various characteristics.
Table 4 gives the GIST scores of age for trial sets in different study phases, study sponsors, and study designs on seven medical conditions, horizontally ordered in the same order as Table 3. For all the conditions, the GIST score of age increases from Phase I to Phase III. This is in accordance with the fact that Phase I trials aim to establish initial safety and efficacy profile of a treatment in a small group of patients, while Phase III trials seek to test the treatment with a large groups of people to confirm its safety and efficacy.

According to the GIST scores, industry-sponsored trials have better population representativeness than NIH-sponsored trials for asthma, sleep disorders, depression, T2DM, and arthritis. In general, randomized trials have a slightly better population representativeness than non-randomized trials except for asthma trials. With regards to primary purpose (study design), treatment and diagnostic trials have better population representativeness than prevention and basic science trials. These results confirmed our Hypothesis #2 that GIST correlates with the population representativeness of a set of related trials.

Table 4. The GIST scores of age for trials in different phases, sponsors, and study designs on seven medical conditions. The number of trials is enclosed by parentheses.

<table>
<thead>
<tr>
<th>Trial Characteristics</th>
<th>Asthma</th>
<th>Sleep disorders</th>
<th>Depression</th>
<th>T2DM</th>
<th>Arthritis</th>
<th>Stroke</th>
<th>Heart attack</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study Phase</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase I</td>
<td>0.53 (179)</td>
<td>0.68 (76)</td>
<td>0.67 (229)</td>
<td>0.60 (368)</td>
<td>0.71 (261)</td>
<td>0.75 (183)</td>
<td>0.79 (60)</td>
</tr>
<tr>
<td>Phase II</td>
<td>0.58 (410)</td>
<td>0.75 (188)</td>
<td>0.72 (408)</td>
<td>0.77 (517)</td>
<td>0.84 (566)</td>
<td>0.84 (319)</td>
<td>0.87 (172)</td>
</tr>
<tr>
<td>Phase III</td>
<td>0.59 (336)</td>
<td>0.80 (193)</td>
<td>0.81 (384)</td>
<td>0.87 (766)</td>
<td>0.86 (582)</td>
<td>0.85 (212)</td>
<td>0.92 (160)</td>
</tr>
<tr>
<td>Phase IV</td>
<td>0.52 (231)</td>
<td>0.69 (136)</td>
<td>0.78 (318)</td>
<td>0.80 (484)</td>
<td>0.83 (404)</td>
<td>0.82 (131)</td>
<td>0.93 (173)</td>
</tr>
<tr>
<td><strong>Sponsor Type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIH</td>
<td>0.31 (27)</td>
<td>0.64 (10)</td>
<td>0.65 (44)</td>
<td>0.57 (35)</td>
<td>0.81 (29)</td>
<td>0.86 (27)</td>
<td>1.00 (2)</td>
</tr>
<tr>
<td>Industry</td>
<td>0.60 (778)</td>
<td>0.80 (287)</td>
<td>0.80 (431)</td>
<td>0.80 (1464)</td>
<td>0.85 (1237)</td>
<td>0.85 (234)</td>
<td>0.92 (145)</td>
</tr>
<tr>
<td>U.S. Fed</td>
<td>0.45 (6)</td>
<td>0.75 (44)</td>
<td>0.93 (49)</td>
<td>0.81 (22)</td>
<td>0.89 (30)</td>
<td>0.93 (44)</td>
<td>1.00 (2)</td>
</tr>
<tr>
<td>Other</td>
<td>0.48 (648)</td>
<td>0.72 (654)</td>
<td>0.73 (1432)</td>
<td>0.73 (1181)</td>
<td>0.78 (894)</td>
<td>0.81 (793)</td>
<td>0.89 (489)</td>
</tr>
<tr>
<td><strong>Study Design - Allocation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomized</td>
<td>0.54 (1217)</td>
<td>0.75 (772)</td>
<td>0.76 (1525)</td>
<td>0.77 (2319)</td>
<td>0.83 (1661)</td>
<td>0.84 (867)</td>
<td>0.90 (554)</td>
</tr>
<tr>
<td>Non-Randomized</td>
<td>0.57 (140)</td>
<td>0.70 (93)</td>
<td>0.71 (201)</td>
<td>0.74 (217)</td>
<td>0.80 (295)</td>
<td>0.79 (104)</td>
<td>0.84 (43)</td>
</tr>
<tr>
<td><strong>Study Design – Primary Purpose</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>0.56 (1077)</td>
<td>0.75 (767)</td>
<td>0.77 (1485)</td>
<td>0.80 (2043)</td>
<td>0.83 (1872)</td>
<td>0.83 (805)</td>
<td>0.89 (484)</td>
</tr>
<tr>
<td>Prevention</td>
<td>0.34 (81)</td>
<td>0.73 (24)</td>
<td>0.54 (139)</td>
<td>0.63 (217)</td>
<td>0.84 (62)</td>
<td>0.80 (151)</td>
<td>0.91 (65)</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>0.56 (57)</td>
<td>0.83 (48)</td>
<td>0.70 (36)</td>
<td>0.77 (40)</td>
<td>0.79 (34)</td>
<td>0.79 (33)</td>
<td>0.95 (42)</td>
</tr>
<tr>
<td>Basic Science</td>
<td>0.53 (78)</td>
<td>0.57 (50)</td>
<td>0.59 (56)</td>
<td>0.64 (162)</td>
<td>0.66 (56)</td>
<td>0.57 (13)</td>
<td>0.54 (10)</td>
</tr>
</tbody>
</table>

Discussion

This study validated the effectiveness of GIST at assessing the population representativeness of related clinical trials. The GIST scores consistently agreed with the expected differences of population representativeness for three population samples across 100 experiments for all three selected eligibility features. We further demonstrated the effectiveness of GIST in comparing population representativeness of trial sets that differ in their characteristics such as disease domains, sponsor types, study phases, and study designs. Among seven medical conditions, asthma trials had the lowest GIST score of age, reflecting the concern in the respiratory medicine research community [20]. Meanwhile, the GIST metric was further validated by the increasing GIST scores of age from Phase I to Phase III across all the seven conditions. Note that GIST can also assess the population representativeness of a single clinical study, as its primary use case of a pharmaceutical company will be to assess the generalizability of a study it is currently designing or even a study that it already has completed.

In this work, we used NHANES to profile the “real-world” patient populations (P1) and weighted “real-world” patient populations (P2). Compared with EHR data, NHANES has several advantages. First, its sophisticated sampling mechanism ensures the population representativeness at the national level. In contrast, EHR data contain mostly diseased patients or patients receiving care and hence may be biased towards certain population subgroups. Second, structured survey data are readily analyzable, whereas EHR data often require preprocessing to address data
quality problems. Therefore, NHANES is more cost-effective than EHR data for lightweight population-based studies. However, due to the limited data in NHANES, it may not be suitable for longitudinal analysis or studies on medical conditions that are not included in the interview questions.

Limitations
This study has limitations. We only included seven medical conditions that have a fair amount of patients (over 1,000) in NHANES. Ideally, more conditions should be analyzed. The self-reported medical conditions in NHANES may have resulted in some misclassification of samples. The GIST metric has intrinsic limitations. First, GIST does not take into account the enrollment value of a study. Currently, ClinicalTrials.gov has only one field for enrollment, which can be planned or actual. Quite a number of trials have not updated the planned enrollment with actual enrollment even after completion. Second, it does not consider the geographic location of the trial, which is one major factor for patient recruitment. Nevertheless, by aggregating many patients and clinical trials, we have minimized the impact of these factors for generating meaningful results in the research community level. Third, the GIST metric cannot reveal the reason behind the population representativeness problem. Visualization such as Figure 3 can serve as a good complement to GIST for assessing the population representativeness of related trials.

The long-term goals of this line of research
As the main purpose of most RCTs is to test the efficacy and safety of a treatment for a certain medical condition in people, it is often required to minimize confounding factors that may potentially affect the results. Therefore, it is a common practice that RCTs usually recruit patients who do not have comorbidities and are not too old or too sick to treat. Instead of enforcing a trial to be generalizable to the broad patient population, the goals of this line of research are (1) to improve the transparency of clinical trial eligibility criteria design biases across multiple studies; (2) to facilitate evidence-based data-driven precision design of clinical trial eligibility criteria [21]; and (3) to address the rising need for patient-centered outcomes research in the clinical trial domain. This information can be provided to clinical trial designers to help them better justify the trade-offs between the internal validity and the external validity when designing a new trial. This information can also help clinical investigators and policy makers efficiently identify population representativeness issues in clinical studies of certain characteristics and take measures accordingly. When applying clinical trial eligibility criteria to observational data, one can compare and contrast the effects observed between eligible and ineligible patients, which may reveal more profound problems in trial design and clinical research in general.

Future work
In the future, we plan to use GIST to identify restrictive features among multiple frequently used eligibility features among clinical trials of a certain medical condition. We will first leverage controlled terminologies such as SNOMED CT to develop structural, semantic, and lexical methods for meaningful aggregation of similar qualitative features. Then we will compute GIST scores to identify stringent features (with relatively low GIST scores). In this paper, we compared the population representativeness of trials on seven medical conditions with respect to age using GIST. However, this method is not efficient when more eligibility features are included in the analysis. Moreover, eligibility features may have inherent correlations. For example, a previous study has reported that impaired fasting glucose generally increases with age for diabetic patients [22]. Meanwhile, eligibility criteria may be operationalized as an interaction of multiple features, e.g., “pregnant female over 40 years old.” In the future, we will investigate how to assess the collective population representativeness using multiple eligibility features simultaneously.

Conclusions
In this work, we used real-world population-level data to validate a novel metric for quantifying the population representativeness of clinical trials. The study results confirmed that the GIST metric is reliable for its purpose. These findings suggested the future potential of a systematic approach for providing prognostic tools to facilitate the clinical trial design process as well as post hoc evaluations to investigate the generalizability of studies already underway or completed. By integrating the real-world experience of patients with the study design attributes of existing clinical trials, researchers designing new clinical studies can improve both the efficiency and generalizability of their designs with this proactive data-driven approach.

Acknowledgments
This study was sponsored by the U.S. National Library of Medicine grant R01LM009886 (PI: Weng) and U.S. National Center for Advancing Translational Science grant UL1 TR000040 (PI: Ginsberg).
References

2. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93.
JuFiT: A Configurable Rule Engine for Filtering and Generating New Multilingual UMLS Terms

Johannes Hellrich, Stefan Schulz, MD, Sven Buechel, Udo Hahn, PhD

1 Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany; 2 Medizinische Universität Graz, Graz, Austria

Abstract

We here describe JuFiT, an easily adjustable rule engine which allows to filter non-natural terms (i.e., ones usually not occurring in running citation texts) from the UMLS metathesaurus and even adds new terms to the UMLS (by rewriting non-natural terms). Unlike previous attempts (with METAMap or CASPER), JuFiT serves multilingual purposes in that it runs for English, Spanish, French, German and Dutch documents, as well – the most prominent European languages in terms of UMLS coverage. We evaluated JuFiT under a variety of experimental conditions and found evidence that it increases annotation quality for English, and most likely also for German and Spanish.

Introduction

The Unified Medical Language System (UMLS) metathesaurus (‘metathesaurus’ will be omitted for brevity) constitutes one of the most important biomedical term repositories, integrating 179 source vocabularies in 21 natural languages. Unfortunately, a large number of UMLS entries are ill-suited for usage besides manual indexing, since they occur in a ‘non-natural’ format, e.g., as terms exhibiting syntactic inversions (such as ‘failure, renal’ instead of ‘renal failure’) or contain semantic meta data, as is the case for terms that originate from SNOMED CT. Such UMLS terms, however, are unlikely to appear in real texts, like clinical narratives or medical literature, and are thus probably worthless or even harmful for advanced NLP applications like information extraction or text mining.

Such non-natural appearances of entries can be removed or even corrected with automatic tools, as prominently demonstrated by the METAMap system. Yet, currently there exists no general-purpose tool for dealing with such non-natural formats for UMLS entries. METAMap acts as a filter (e.g., by correcting syntactic inversions or generating variants without possessives) when used to process text. However, its filtering component is not available as a stand-alone tool and thus inapplicable for many applications. In addition, METAMap is known to be rather slow. The CASPER tool was developed as a stand-alone solution for filtering and correcting UMLS entries, yet, just as METAMap, focuses on English only. This is troublesome in times when multilingualism is considered an important issue within the health care community due to ever-increasing economic globalization, international tourism and immigration streams. Improved term quality could not only improve biomedical NLP for non-English languages but also yield benefits for machine translation systems and the internationalization of the terminological infrastructure of the life sciences as evidenced by the UMLS, which is still very much English-centric (cf. Table 1).

Table 1 Absolute and relative number of raw UMLS term strings (SUs) for the languages covered in our experiments.

<table>
<thead>
<tr>
<th>Language</th>
<th>Terms</th>
<th>Percentage of English</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>7,377,229</td>
<td>100.0%</td>
</tr>
<tr>
<td>Spanish</td>
<td>1,095,842</td>
<td>14.9%</td>
</tr>
<tr>
<td>Dutch</td>
<td>192,521</td>
<td>2.6%</td>
</tr>
<tr>
<td>French</td>
<td>178,861</td>
<td>2.4%</td>
</tr>
<tr>
<td>German</td>
<td>163,277</td>
<td>2.2%</td>
</tr>
</tbody>
</table>

Notes:

a) [http://www.ihtsdo.org/snomed-ct](http://www.ihtsdo.org/snomed-ct)
c) [http://www.biosemantics.org/index.php/software/casper](http://www.biosemantics.org/index.php/software/casper)
d) [https://github.com/JULIELab/jufit](https://github.com/JULIELab/jufit)
e) [http://www.biosemantics.org/index.php/software/casper](http://www.biosemantics.org/index.php/software/casper)
f) We here included SNOMED, MEDCIN, the National Drug File – Reference Terminology, the Foundational Model of Anatomy Ontology, the
We alleviate this lack of a stand-alone tool suited for multilingual environments by providing the Jena UMLS Filter Tool (JuFiT), a rule-based open-source tool.\(^d\) JuFiT uses rules similar to those in MetaMap and CASPER, but extends these by adding support for multiple languages and their language-specific phenomena, e.g., German compounds. We provide rules for English, Spanish, French, German, and Dutch, yet the declarative implementation of these rules makes it possible to add support for further languages without changing JuFiT’s source code.

Using rule-based systems to transform biomedical terms is, by no means, a new idea. It originates from the work of McCray and colleagues for the English language carried out at the U.S. National Library of Medicine.\(^{10-12}\) Such transformations were previously shown to be useful for mapping terminologies\(^{13}\) and for increasing system performance in retrieval experiments.\(^{14}\) Filtering UMLS entries could also lead to consistent improvements in terminology-assisted machine translation, where results are quite mixed at the moment.\(^{7,15}\)

Our rules were evaluated both manually and automatically, first, by looking at the number of dictionary entries as influenced by each rule, second, by using these dictionaries to annotate multilingual texts and counting the number of matches and, third, by taking a sample of terms only annotated or no longer annotated thanks to our rules. The multilingual texts were taken from the cleaned EMEA corpus\(^6\) portion of the MANTRA challenge material\(^6\) and can be assumed to be direct translations of each other over all languages. We could demonstrate a pronounced increase in annotation quality for English and, arguably, for German and Spanish, as well as a moderate increase (~2\%) in annotation coverage for French and Dutch thanks to the operation of JuFiT.

**Methods**

All experiments were conducted based on an installation of the 2014AB UMLS (processing the MRCONSO.RFF and MRSTY.RFF files) with default settings and all sources. We applied both single rules and combinations thereof using appropriate JuFiT configuration files to produce dictionaries digestible by a LingPipe\(^e\) based gazetteer, suppressing terms marked as suppressible in the UMLS (the latter was also done for the baseline). Filtering was applied to prevent rules from producing duplicate strings, i.e., ones that were already contained in the UMLS for a term of the language under scrutiny. Whereas JuFiT is capable of preserving the original cases (helpful for blacklist-based rules), the gazetteer was configured to ignore them in order to increase recall (many UMLS terms reveal irregular capitalization).

Two types of rules can be distinguished, viz. deletion rules (removing terms) and rewrite rules (generating variants, while preserving the original term, with the exception of RewriteSemanticType). The basic principles underlying these rules are well-known in the literature\(^{2,4,10-12}\) yet, we differ from prior works in terms of increasing lexical coverage and handling multilingual phenomena.

**Deletion rules:**

- DeleteIfContainsAtSign: Deletion of terms containing “@”, such as “CASTOR OIL@@MISCELL@OIL”.
- DeleteIfContainsResiduals: Deletion of terms marked as unspecific through warning strings, e.g., “not elsewhere classified” or “other” for English, “no clasificados” (and other inflected forms) for Spanish. This rule combines the “Any classification”, “Any underspecification” and “Miscellaneous” rules from Hettne et al.\(^4\)
- DeleteIfContainsDosage: Deletion of terms containing dosages and other measurement information, covering, e.g., “5 ml”, “Milligram/Kilogram“, “per day” or “2 %“.
- DeleteIfContainsECNumber: Deletion of terms containing enzyme classifications, identified as terms containing “EC” followed by at least one digit and zero up to three combinations of a dot followed by a single or multiple digits.
- DeleteShortToken: Deletion of terms that do not contain at least one token with two or more letters, which is not a Roman numeral (between I and VIII) or a stop word (e.g., German “die” [English: “the”]). This rule removes terms such as “L.Q.; 20-34”.

\(^d\) https://github.com/JULIELab/jufit
\(^e\) http://alias-i.com/lingpipe/
Rewrite rules:

- RewriteApostropheS: Adds a variant without “’s” at the end of words. This results in variants without possessives in English, e.g., “Abelson virus” for “Abelson’s virus”, and plurals in Dutch, e.g., “Getto” [English: “ghetto”] for “Getto’s” [English: “ghettos”]. This rule was only used for English and Dutch.

- RewriteShortFormLongForm: Adds separate variants for terms containing both a long and a short form by using Schwartz and Hearst’s algorithm, resulting in new term variants like “child ADD” and “child attention deficit disorder” for “child attention deficit disorder (child ADD)”.

- RewriteSyntacticInversion: Adds a variant of terms in which syntactic inversions are reverted, e.g., “Adult Human” for “Human, Adult”. We provide special support for compounding according to German and Dutch orthographic conventions, i.e. “Berufsunfälle” [English: “work accidents”] is added for “Unfälle, Berufs-” instead of (orthographically false) “Berufs-Unfälle”.

- RewriteSemanticType: Replaces terms containing a semantic type in rounded, angular or square brackets by a variant where the brackets and their contents are removed. Separate lists are used to identify types contained in rounded, angular and square brackets for each language. These types are not identical with the UMLS semantic types (as in CASPER⁷), yet were manually extracted from frequent terms with brackets in several resources.⁷ We also added some rules to all languages, e.g. “SMQ” (for “Standardised MedDRA Queries”),⁸ as well as the language-specific names for the “Kiel Classification” (appearing in most terms describing non-Hodgkin lymphoma),⁹ after searching the UMLS for language-specific bracket content in German, French and Dutch. This rule yields terms like “injection of steroid into spinal canal” as a replacement for “injection of steroid into spinal canal (procedure)”, a typical example of a SNOMED CT Fully Specified Name.

Rule implementation is as generic as possible, language-specific variants or word lists, as well as the active rules themselves and their sequence order are configured via a file, allowing for easy adaptation to additional languages. These rules were organized in three different rule sequences, i.e. all rewrite rules, all deletion rules and a balanced sequence (following Hettne et al. who combined both types of rules). Rules can modify term variants generated by other rules in order to ensure that no new term would conflict with the constraints imposed by our deletion rules.

Evaluation

The evaluation was performed under three conditions, first, by looking at the number of dictionary entries produced or removed by each rule, second, by using these dictionaries to annotate EMA files and counting the number of matches and, third, by taking a random sample of terms only annotated or no longer annotated given our rules. The manual evaluation was performed on samples for English, Spanish, German and French for the balanced condition. As evident from Table 2, the balanced rule sequence resulted in an increase of dictionary entries for German and Dutch only, whereas English and Spanish lost 12.8% and 36.9% of their respective entries. One of the main reasons for this surprising effect is due to the application of the RewriteSemanticType rule—many of the terms affected by this rule exist both with and without the interfering semantic type metadata (removed by the rule). This is typical for SNOMED CT, which constitutes a substantial part of the UMLS’ English and Spanish contents. These terms are effectively removed by the rule, as no rule may produce a new variant which is indistinguishable from existing terms. This effect is far less pronounced in Hettne et al.’s study,⁴ due to their questionable semantic type selection. The highest number of additional term variants for all rules was provided by the RewriteSyntacticInversion rule, which is consistent with numbers reported by Hettne et al.⁴ (their data show an 8.6% increase, whereas we found a

---

⁷ We here included SNOMED, MEDCIN, the National Drug File – Reference Terminology, the Foundational Model of Anatomy Ontology, the NCBI Taxonomy, and the UMLS Metathesaurus – covering English and (via SNOMED) Spanish.

⁸ http://www.meddra.org/standardised-meddra-queries

⁹ http://www.morbus-hodgkin.de/infoserv/kiel.htm

⁴ For the balanced condition, we executed the rules in the following sequence: DeleteIfContainsDosage, DeleteIfContainsAtSign, DeleteIfContainsECNumber, DeleteIfContainsResidual, DeleteShortToken, RewriteSemanticType, RewriteAngularBrackets, RewriteShortFormLongForm, RewriteApostropheS (English/Dutch only), DeleteIfContainsDosage, DeleteIfContainsAtSign, DeleteIfContainsECNumber, DeleteIfContainsResidual, DeleteShortToken; the deletion rules were applied twice to cover newly added variants as well.
10.9% increase, for English terms). Deletion rules had only minor effects, the strongest were seen for DeleteIfContainsDosage for Spanish and English, again consistent with results from Hettne et al.\(^4\)

Table 2. Number of terms in the gazetteer file produced by JuFiT per language for all tested rules and rule combinations. ‘Balanced’ combines the deletion and rewrite rules. Baseline numbers differ from the size numbers reported in Table 1 since suppressible terms, as well as identical term strings for the same CUI, were removed.

<table>
<thead>
<tr>
<th></th>
<th>German</th>
<th>Spanish</th>
<th>English</th>
<th>French</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>148,648</td>
<td>813,917</td>
<td>5,774,479</td>
<td>163,681</td>
<td>178,055</td>
</tr>
<tr>
<td>Balanced</td>
<td>152,733</td>
<td>513,205</td>
<td>5,036,283</td>
<td>161,859</td>
<td>195,752</td>
</tr>
<tr>
<td>All deletion rules</td>
<td>140,296</td>
<td>790,492</td>
<td>5,169,712</td>
<td>158,908</td>
<td>170,939</td>
</tr>
<tr>
<td>All rewrite rules</td>
<td>162,990</td>
<td>530,366</td>
<td>5,822,297</td>
<td>163,053</td>
<td>178,026</td>
</tr>
<tr>
<td>DeleteIfContainsAtSign</td>
<td>148,648</td>
<td>813,811</td>
<td>5,596,248</td>
<td>163,681</td>
<td>178,055</td>
</tr>
<tr>
<td>DeleteIfContainsDosage</td>
<td>148,393</td>
<td>796,189</td>
<td>5,462,101</td>
<td>162,940</td>
<td>177,830</td>
</tr>
<tr>
<td>DeleteIfContainsECNumber</td>
<td>148,648</td>
<td>813,917</td>
<td>5,773,295</td>
<td>163,681</td>
<td>178,055</td>
</tr>
<tr>
<td>DeleteIfContainsResidual</td>
<td>140,595</td>
<td>808,778</td>
<td>5,681,281</td>
<td>159,798</td>
<td>171,175</td>
</tr>
<tr>
<td>DeleteShortToken</td>
<td>148,592</td>
<td>813,291</td>
<td>5,759,449</td>
<td>163,521</td>
<td>178,026</td>
</tr>
<tr>
<td>RewriteApostropheS</td>
<td>–</td>
<td>–</td>
<td>5,800,253</td>
<td>–</td>
<td>178,437</td>
</tr>
<tr>
<td>RewriteSemanticType</td>
<td>148,525</td>
<td>504,627</td>
<td>5,185,218</td>
<td>163,617</td>
<td>178,006</td>
</tr>
<tr>
<td>RewriteShortFormLongForm</td>
<td>148,711</td>
<td>814,849</td>
<td>5,778,934</td>
<td>165,013</td>
<td>178,154</td>
</tr>
<tr>
<td>RewriteSyntacticInversion</td>
<td>269,212</td>
<td>854,965</td>
<td>6,403,773</td>
<td>165,952</td>
<td>205,104</td>
</tr>
</tbody>
</table>

Table 3 depicts the number of entity occurrences in the EMEA corpus that were annotated with gazetteer dictionaries, filtered by different rules and rule sequences. The balanced rule sequence caused little deviation from the absolute number of German annotations, while for French we observed an increase of 2.1% and 2.4% for Dutch. By contrast, both English and Spanish lost several tens of thousands annotations, 3.2% and 1.6%, respectively. This drop is inconsistent with results reported by Hettne et al.,\(^4\) which showed a 2.8% increase.

Table 3. Number of annotations (per language) produced by LINGPipe after preprocessing by JuFiT in the EMEA corpus for all tested rules and rule combinations. ‘Balanced’ combines the deletion and rewrite rules.

<table>
<thead>
<tr>
<th></th>
<th>German</th>
<th>Spanish</th>
<th>English</th>
<th>French</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>221,639</td>
<td>646,118</td>
<td>1,786,834</td>
<td>309,460</td>
<td>214,905</td>
</tr>
<tr>
<td>Balanced</td>
<td>222,801</td>
<td>636,073</td>
<td>1,730,465</td>
<td>316,061</td>
<td>220,094</td>
</tr>
<tr>
<td>All deletion rules</td>
<td>221,619</td>
<td>635,727</td>
<td>1,730,277</td>
<td>309,102</td>
<td>214,925</td>
</tr>
<tr>
<td>All rewrite rules</td>
<td>222,821</td>
<td>646,464</td>
<td>1,787,352</td>
<td>319,217</td>
<td>220,848</td>
</tr>
<tr>
<td>DeleteIfContainsAtSign</td>
<td>221,639</td>
<td>646,118</td>
<td>1,786,835</td>
<td>309,460</td>
<td>214,905</td>
</tr>
<tr>
<td>DeleteIfContainsDosage</td>
<td>221,633</td>
<td>640,904</td>
<td>1,759,785</td>
<td>309,421</td>
<td>214,921</td>
</tr>
<tr>
<td>DeleteIfContainsECNumber</td>
<td>221,639</td>
<td>646,118</td>
<td>1,786,834</td>
<td>309,460</td>
<td>214,905</td>
</tr>
<tr>
<td>DeleteIfContainsResidual</td>
<td>221,640</td>
<td>646,114</td>
<td>1,786,941</td>
<td>309,464</td>
<td>214,910</td>
</tr>
<tr>
<td>DeleteShortToken</td>
<td>221,624</td>
<td>640,604</td>
<td>1,744,964</td>
<td>309,137</td>
<td>214,904</td>
</tr>
<tr>
<td>RewriteApostropheS</td>
<td>–</td>
<td>–</td>
<td>1,787,038</td>
<td>–</td>
<td>214,905</td>
</tr>
<tr>
<td>RewriteSemanticType</td>
<td>222,600</td>
<td>646,355</td>
<td>1,786,800</td>
<td>309,620</td>
<td>215,180</td>
</tr>
<tr>
<td>RewriteShortFormLongForm</td>
<td>221,718</td>
<td>751,769</td>
<td>1,788,105</td>
<td>319,494</td>
<td>220,455</td>
</tr>
<tr>
<td>RewriteSyntacticInversion</td>
<td>221,781</td>
<td>646,138</td>
<td>1,786,211</td>
<td>309,460</td>
<td>215,022</td>
</tr>
</tbody>
</table>

Table 3
To better assess the differences between JuFiT and CASPER we used the latter to processes the UMLS version of the EMEA corpus (with LINGPIPE) used in our JuFiT experiments for English. The results of this experiment are listed in Table 4. CASPER’s rules can mostly be mapped 1:1 to JuFiT rules (only DeleteIfContainsResidual is a combination of CASPER’s “Any classification”, “Any underspecification” and “Miscellaneous” rules). Hence, we use JuFiT’s rule names for ease of comparison. The number of terms is counted after the removal of multiply identical terms for the same CUI. The number of annotations produced by a ‘balanced’ system is quite similar for both JuFiT and CASPER, thus the differences in UMLS versions or corpora must be the source of the aforementioned drop in comparison with the baseline. For the three tested rule combinations, JuFiT produces fewer terms, yet leads to slightly more annotations than CASPER. For single rules, the number of annotations differ only slightly between both systems, except for the DeleteIfContainsDosage rule where JuFiT’s implementation is more general and thus more effective. The number of terms produced by applying single rules differs strongly for two rules—JuFiT’s RewriteSemanticType rule produces only 90% of the terms generated by CASPER, which can probably be attributed to our finer grained list of semantic types (cf. rule description above for details). Meanwhile, for RewriteSyntacticInversion JuFiT produces 5% more terms, which may be due to our implementation that includes no checks for prepositions or conjunctions, whereas CASPER’s implementation does. These explanations are educated guesses, since CASPER is not available open source, in contrast to JuFiT.

We refrained from comparing JuFiT with METAMap because they markedly differ in design. METAMap is a full-blown lexical processing engine, whereas JuFiT is only a lexical pre-processor without, e.g., disambiguation or tagging functionality. Hence, a comparison of both would require embedding JuFiT into a more comprehensive text analysis pipeline. Still, such a comparison is badly needed and a topic of future research.

Table 4. Number of terms and annotations achieved by using CASPER (and LINGPIPE) on the 2014AB UMLS and annotating EMEA with the filtered terminology (English only due to CASPER’s monolingual nature).

<table>
<thead>
<tr>
<th>Rule</th>
<th>Terms</th>
<th>Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balanced</td>
<td>5,540,257</td>
<td>1,728,158</td>
</tr>
<tr>
<td>All deletion rules</td>
<td>5,220,665</td>
<td>1,728,490</td>
</tr>
<tr>
<td>All rewrite rules</td>
<td>6,136,840</td>
<td>1,786,423</td>
</tr>
<tr>
<td>DeleteIfContainsAtSign</td>
<td>5,588,309</td>
<td>1,786,794</td>
</tr>
<tr>
<td>DeleteIfContainsDosage</td>
<td>5,529,162</td>
<td>1,787,796</td>
</tr>
<tr>
<td>DeleteIfContainsECNumber</td>
<td>5,765,546</td>
<td>1,786,793</td>
</tr>
<tr>
<td>DeleteIfContainsResidual</td>
<td>5,641,098</td>
<td>1,772,832</td>
</tr>
<tr>
<td>DeleteShortToken</td>
<td>5,764,069</td>
<td>1,742,281</td>
</tr>
<tr>
<td>RewriteApostropheS</td>
<td>5,791,568</td>
<td>1,786,764</td>
</tr>
<tr>
<td>RewriteSemanticType</td>
<td>5,766,859</td>
<td>1,786,793</td>
</tr>
<tr>
<td>RewriteShortFormLongForm</td>
<td>5,766,943</td>
<td>1,787,064</td>
</tr>
<tr>
<td>RewriteSyntacticInversion</td>
<td>6,110,934</td>
<td>1,786,181</td>
</tr>
</tbody>
</table>

Finally, we performed a manual review by comparing the terms annotated in the baseline with those annotated after balanced preprocessing. We sorted these terms in three categories, those annotated only in the baseline, those annotated only by JuFiT and those annotated differently by JuFiT, i.e. the baseline contained a term, yet a longer and overlapping term was found, e.g., “Stage 5 Chronic Kidney Disease” instead of the more general “Chronic Kidney Disease”. This analysis was performed for English, Spanish, German and French by taking the 10 most frequent changes for each language plus 50 randomly selected ones into consideration. These samples were annotated as positively, moderately or negatively effecting annotation quality, with results shown in Table 5. Removal of all annotations for measurements was intended and thus rated positively. According to our analysis, JuFiT was clearly helpful for English, and most likely also for German and Spanish, yet changes for French were about equally positive and negative. The differences between languages could point both towards additional efforts necessary for some languages (especially French), yet can also be explained by large, language-specific differences in terms of source vocabularies in the UMLS.
Table 5. Number of annotations positively, moderately or negatively affected when comparing baseline annotations with annotations generated after application of JuFiT’s balanced rule sequence; results are based on a sample of 60 changes per language.

<table>
<thead>
<tr>
<th></th>
<th>Positive effect</th>
<th>Moderate effect</th>
<th>Negative effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Spanish</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>German</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>French</td>
<td>7</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Random 50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>46</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Spanish</td>
<td>32</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>German</td>
<td>44</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>French</td>
<td>24</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>56</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Spanish</td>
<td>41</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>German</td>
<td>49</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>French</td>
<td>31</td>
<td>1</td>
<td>28</td>
</tr>
</tbody>
</table>

Conclusions

While the UMLS is of central importance for numerous clinical and biomedical NLP applications, the structure of some of its entries, containing e.g., syntactic inversions (‘failure, renal’ instead of ‘renal failure’, the latter being recoverable by the RewriteSyntacticInversion rule), hampers these applications. We provide JuFiT, an open source tool for filtering and improving such entries, which extends prior solutions (such as METAmap and CASPER) by adding analytic capabilities for several non-English languages. The impact of JuFiT was tested in three kinds of evaluation experiments, one manual and two automatic ones. We could demonstrate a clear increase in annotation quality for English and arguably German and Spanish, as well as a moderate increase (~2%) in annotation coverage for French and Dutch with JuFiT’s balanced rule sequence.

Future work could be directed in several directions, e.g., adding additional rules for further phenomena, e.g., TNM classifications, improving existing rules to avoid false positives or increasing the number of languages for which specific rules exist. Furthermore, a comparison between METAmap and an analysis pipeline with JuFiT embedded in it is on the agenda. As mentioned before, one could also try to use the filtered UMLS for other applications besides information extraction, e.g., as a resource for machine translation.

References

Mining Twitter as a First Step toward Assessing the Adequacy of Gender Identification Terms on Intake Forms

Amanda Hicks, PhD\textsuperscript{1}, William R. Hogan, MD, MS\textsuperscript{1}, Michael Rutherford, MS\textsuperscript{2}, Bradley Malin, PhD\textsuperscript{3}, Mengjun Xie, PhD\textsuperscript{4}, Christiane Fellbaum, PhD\textsuperscript{5}, Zhijun Yin, BE\textsuperscript{5}, Daniel Fabbri, PhD\textsuperscript{3}, Josh Hanna, MS\textsuperscript{1}, Jiang Bian, PhD, MS\textsuperscript{1}
\textsuperscript{1}University of Florida, Gainesville, FL; \textsuperscript{2}University of Arkansas for Medical Sciences, Little Rock, AR; \textsuperscript{3}Vanderbilt University, Nashville, TN; \textsuperscript{4}University of Arkansas at Little Rock, Little Rock, AR; \textsuperscript{5}Princeton University, Princeton, NJ

Abstract
The Institute of Medicine (IOM) recommends that health care providers collect data on gender identity. If these data are to be useful, they should utilize terms that characterize gender identity in a manner that is 1) sensitive to transgender and gender non-binary individuals (trans* people) and 2) semantically structured to render associated data meaningful to the health care professionals. We developed a set of tools and approaches for analyzing Twitter data as a basis for generating hypotheses on language used to identify gender and discuss gender-related issues across regions and population groups. We offer sample hypotheses regarding regional variations in the usage of certain terms such as ‘genderqueer’, ‘genderfluid’, and ‘neutrois’ and their usefulness as terms on intake forms. While these hypotheses cannot be directly validated with Twitter data alone, our data and tools help to formulate testable hypotheses and design future studies regarding the adequacy of gender identification terms on intake forms.

Introduction
The LGBT community is subject to a variety of health disparities. This is a result of a lack of meaningful data on LGBT populations as well as a lack of training and resources for clinicians to provide culturally competent care. Recent Institute of Medicine (IOM) recommendations to address these health disparities include (1) gathering data on sexual orientation and gender identity in Electronic Health Records (EHR) as part of the meaningful use objectives in EHRs, (2) developing standardization of sexual orientation and gender identity measures to facilitate synthesizing scientific knowledge about the health of sexual and gender minorities, and (3) supporting research to develop innovative methods of conducting research with small populations and to determine the best ways to collect information on LGBT minorities.\textsuperscript{1}

While the IOM notes that data collection would be aided by standardized measures for sexual orientation and gender identity, their report also emphasizes that defining sexual orientation and gender nonconformity is a challenge since these are multifaceted concepts. The use of terminology that is familiar to the participant has been shown to improve response rates.\textsuperscript{1,2} However, based on the limited research available, there is some evidence\textsuperscript{3,5} to suggest that consumer vocabulary for self-identifying gender and sexual orientation varies by community. There is clear evidence of lexical variation associated with geography in linguistics studies.\textsuperscript{6,8} Also, through discussions with members of the trans* community and health care providers at LGBT clinics across the country, we have learned that new terms are frequently being coined to describe gender identity and that the connotations of existing terms may vary by community.

There is documented variation of terms to describe sexual orientation across communities.\textsuperscript{9} There is also variation in the meaning of terms between individuals who consider themselves part of the sexual minority (e.g., lesbian, gay, or bisexual) and those who do not (e.g., straight or heterosexual).\textsuperscript{10} For example, self-identifying members of a sexual minority use ‘lesbian’ to refer to women who are primarily attracted to other women, but others tend to use ‘lesbian’ more broadly to refer to a woman who has experienced any sexual attraction or sexual activity with another woman.\textsuperscript{10} This raises the question of whether there is similar variation in the meanings of terms used to describe transgender identity. However, data addressing variations of gender identity terms and their meanings is lacking. This is significant for the development of good intake forms; if there is significant lexico-semantic variation of gender identity terms, then a single, universal standard intake form may result in a lower response rate than intake forms that are community specific.
A number of organizations have attempted to address the question of how to ask patients about their gender identity. A summary of these approaches can be found in the GenIUSS Report by the Williams Institute. The most promising is a two-step format recommended by the UCSF Center of Excellence for Transgender Health. First patients are asked about their gender identity and then their sex assigned at birth. However, this research addresses the form of the question, not the specific items used to present gender-identity options that ought to be available on the form. The language used in the gender identity question varies across forms from different healthcare organizations. Table 1 contains the choices from the sample forms of three institutions: 1) Fenway Health, 2) UCSF, and 3) the Williams Institute. Although the two-step format has been field tested in Michigan by the Fenway Institute, it is not clear to what extent the terms on these forms represent the identity terms used by transgendered, non-binary, and/or gender-variant people (trans*) across the United States. For brevity we refer to transgendered, gender non-binary and gender-variant people by the term ‘trans*’. The result is that there are still outstanding questions regarding which terms are optimal for intake forms and whether a single, universal standard terminology will suffice for all trans* communities.

<table>
<thead>
<tr>
<th>Fenway Health Intake Form</th>
<th>UCSF Center of Excellence Sample form</th>
<th>GenIUSS Sample Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Male</td>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
<td>Female</td>
<td>Female</td>
</tr>
<tr>
<td>Genderqueer or not exclusively male or female</td>
<td>Transgender Male/Transman/FTM Transgender Female/Transwoman/MTF Genderqueer Additional category (please specify) Decline to Answer</td>
<td>Trans Male/Trans Man Trans Female/ Trans Woman Genderqueer Different Identity (please state)</td>
</tr>
</tbody>
</table>

Table 1. Gender identity terms found on various intake forms.

User generated content on social media, such as Twitter, is a valuable resource because it can provide a source for gleaning information about people’s daily life to answer scientific questions. We believe this source can produce a data set that can contribute to the IOM priority area to study social influences on LGBT health and to the IOM recommendation to develop innovative methods for conducting research on small populations. Mining social networking resources produces data sets that can be used to investigate social influences of health concerns among transgender persons.

Our goal is to build a data set and visualization tools that can be used as a basis to generate hypotheses for further testing to guide the development of gender identity questions on intake forms. Our process for building these tools was as follows. We first examined which terms are currently used to describe transgender identity on Twitter. Based on existing research on linguistic variation in social media, we hypothesize that the usage of gender identification terms varies by geographical region. Then we geotagged the tweets by US state, classified tweets as authored by self-identifying transgender users, and created a co-occurrence network and term frequency counts to support hypothesis generation with data visualization tools. These co-occurrence counts and frequency counts will form the basis of distributional similarity metrics in future research to help determine a) whether different terms are synonyms; and b) whether some terms are polysemous, i.e., carry multiple distinct meanings. By ‘self-identifying’ we refer to people to have stated that they have a trans* identity in some context through their tweets.

Our approach is consistent with the intersectional perspective recommended by the IOM. The intersectional approach considers sub-populations of the LGBT community based on several orthogonal factors, such as ethnicity and geographical region. Furthermore, the resulting data set can be used to address demographic research, social influences on health, and transgender specific health needs — three of the five priority research areas.

Another goal of this paper is to establish a set of best practices for dealing with social media for extracting useful biomedical knowledge, which can help produce data on small populations through unfettered access to such a “Big Data” source (over 500 million tweets per day).
Methods

The general idea underlying our approach is to identify tweets that are relevant to the discussion of trans* related issues, and then examine the variations in language used for gender identification by different communities, that is, by population (trans* people vs. the general public) and by geographical location (U.S. states). The analysis workflow consists of five main steps, as depicted in Figure 1: 1) collect tweets that are potentially related to discussions about gender identification; 2) preprocess and geotag tweets with their corresponding U.S. state; 3) build supervised classification models based on textual features in the tweets to a) filter out irrelevant tweets and b) find people who are self-identified as trans*; 4) collect relevant (both self-identifying trans* users and users in the general public who discussed trans* related issues) users’ Twitter timelines which consists of all of their tweets in chronological order; and 5) compare the usage of gender identification terms by geographical locations (i.e., by U.S. states) and by population groups (trans* people vs. the general public). Some of the search terms are ambiguous and their meanings are context dependent. For example, the tweet “That Hot Pocket is full of trans fats” is not related to discussions of gender identification even through it contains the keyword. To account for this observation, we engineered a binary classifier to determine the likelihood that a tweet is relevant to the discussion of gender identification and remove those that are unlikely to be irrelevant from the corpus in step 3. We also leverage a number of visualization techniques to provide straightforward and easy-to-understand visual representations – word clouds, co-occurrence matrices, and network graphs – to substantiate our findings. In the following sections, we describe each step and the basic procedures in further detail.

![Figure 1](image.png)

**Figure 1.** The general analysis workflow consists of five steps: 1) collect relevant tweets using the Twitter search API with a search term list; 2) preprocess the collected data to filter out non-English tweets and geotag based on user profiles; 3) build classification models to identify relevant tweets and Twitter users; 4) collect relevant users’ Twitter timelines; and 5) analyze the usage of keyterms through comparing term frequencies and co-occurrences.

**a) Data collection through the Twitter search API**

We developed a set of Python scripts leveraging the twython\(^{15}\) library for accessing the Twitter APIs. We designed our Python crawler, `tweetfilter\(^{16}\)`, to handle various potential runtime exceptions (e.g., the crawler will recover from a system failure automatically and pause collection when it reaches the Twitter API rate limits\(^{17}\)) and distribute the workload across multiple Amazon EC2 instances. The data collection process began with a list of keywords (i.e., search terms) mainly related to gender identification such as ‘transwomen’, ‘genderqueer’, and ‘transmasculine’. We have also included a number of other keywords that could indicate relevance of the tweets to trans* discussions such as ‘testosterone’ (often used as part of the hormone replacement therapy for transgender individuals) and ‘gender reassignment surgery’. To ensure coverage, we considered the base forms of these terms as well as their spelling variations, such as ‘transwomen’, ‘trans-women’, and ‘trans women’.

613
Additionally, we found that a number of hashtags (i.e., patterns that start with ‘#’ to mark topics in a tweet and often used by Twitter users to categorize the messages), such as ‘#iamnonbinary’ and ‘#iamtrans’, are good search terms with a low false positive rate for identifying tweets relevant to our study. To develop a list of search terms, we started with ‘transgender’ and ‘trans’ as seed terms which we used as search terms on Twitter and manually compiled a list of co-occurring terms that are in the domain of trans* gender-identification. We next iterated this process until we were no longer accumulating new terms. Then we manually examined the collected tweets to determine the quality of these terms as search terms. Through an iterative process, we removed terms where the majority of the returned tweets were false positive and added new relevant keywords that discovered in the collected tweets.

b) Data preprocessing and geotagging

We preprocessed the collected data to eliminate tweets that 1) were not written in English or 2) those for which we could not determine the geographical location of the user. For language detection, we leveraged the Twitter API metadata directly, which includes a ‘lang’ attribute specifying the language that the tweet was written in. For geotagging, we extracted the ‘location’ field, part of a user’s profile, and attempted to assign a U.S. state to each tweet accordingly. Specifically, we searched each location field for a number of lexical patterns indicating the location of the user such as the name of a state (e.g., Arkansas or Florida), or a city name in combination with a state name or state abbreviation in various possible formats (e.g., “——, fl” or “——, florida” or “——, fl, usa”). Self-reported locations are often non-referring terms (e.g., “wonder land” or “up in the sky”), but strict patterns produced good matches and helped to reduce the number of false positives.

Notably Twitter also provides the ability to attach geocodes (i.e., latitude and longitude) to a user’s profile and to each tweet. Yet, since geolocation needs to be enabled explicitly by the user as well as requires the user to have a device that is capable of capturing geocodes (e.g., a mobile phone with GPS turned on), very few tweets we have collected have this information. This is consistent with findings from previous studies. If the ‘location’ field was missing in a user’s profile, but the ‘geo’ attribute was available, we attempted to resolve the location of the user through reverse geocoding via the publicly available GeoNames geographical database. In Twitter, geocoding can be either at the user-level or at individual tweet-level. However, we did not use the geocodes attached to each individual tweet since it is possible that a user was traveling away from their home state, in which case the geocodes attached to the tweets would be different from those on their profile. For our study, we geotagged the tweets based on where the user is from, not where the user is traveling temporarily. However, we do consider the scenario where a user permanently moved from one state to another reflected as a change in the ‘location’ field of a user’s profile.

We have also made a number of other efforts to clean up the tweets including: 1) fixing Unicode text using ftfy; 2) removing mentions (i.e., indicating conversations in a tweet, starts with ‘@’ followed by a username); and 3) eliminating hyperlinks. However, we did retain hashtags as they indicate topics and categories of the tweets and may contribute to the vocabulary of trans* related discussions.

c) Classification models for finding relevant tweets and Twitter users

Even though a tweet contains one or more of the search keywords, the tweet may not be relevant to our study due to the ambiguity of the search terms. The meanings of many search terms are context dependent. For example, the term ‘trans’ could also mean “trans fat” or “transmission”, depending on the context of the sentence. Since we are interested only in tweets where ‘trans’ means “transgender”, we built a binary classifier to distinguish tweets that are relevant vs. irrelevant to the discussion of gender-related issues. Further, we want to examine whether there are any differences in the terminology used across trans* communities. Thus, we built a second binary classifier to discover people who are self-identified as trans*.

The mechanisms of both classifiers are essentially the same. We first converted each tweet into a feature vector using the term frequency-inverse document frequency (tf-idf) scheme and then trained the classifiers using a random forest. We manually annotated 6,058 tweets to obtain a training sample. All tweets were read by three people and each tweet was assigned one of three labels: ‘irrelevant’ (661 tweets), ‘relevant but NOT self-identifying’ (4,619 tweets), and ‘relevant AND self-identifying’ (778 tweets). When disagreements between the three annotators occurred, we used the majority rule to determine the final label. Although the three labels are mutually exclusive in the sense that only one label is assigned to each tweet, self-identifying tweets are inherently relevant tweets. Therefore, in building the disambiguation classifier, we treated relevant tweets (both self-identifying and not self-identifying) as positive samples and irrelevant tweets as negative samples. In building the second classifier to identify the trans* population, we treated self-identifying tweets as positive and the remainder of the
relevant tweets as negative. We followed standard machine learning best practices (e.g., use 10-fold cross-validation to find the best model parameters—the number of trees in the forest for the random forest model, and for both classifiers the best parameters are 110) to ensure these classifiers are of high quality. The prediction accuracy for finding irrelevant tweets is 97.4% (precision: 0.970; recall: 0.766), and the accuracy for identifying trans* people is 87.8% (precision: 0.741; recall: 0.261).

d) Collect relevant users’ Twitter timelines

Further, we expanded our corpus to include all the tweets posted by the users who were classified as trans*. The motivation for collecting relevant users’ Twitter timelines is two-fold. First, the Twitter search API only returns recent tweets (Twitter does not release the details of their search algorithm, so we do not know exactly how many days of data will be returned prior to the day an inquiry is submitted, but an analysis of our data set suggests it is around 14 days.), and it is important to recognize that a user could have posted discussions related to trans* issues beyond the search limit. Second, our list of search terms does not contain all of the keyterms of interest such that a user could have posted discussions that contain one or more terms that are not search terms. Our search term list is rather restrictive, and does not contain all the gender identification terms that we are interested in to eliminate too many false positives. We removed a term from the search term list when the majority of the tweets it returned are irrelevant. For example, we found that “fm” (“female-to-male”, but could also mean “first time mom”) performed extremely poor. A user’s Twitter timeline can be collected using Twitter’s ‘statuses/user_timeline’ API. However, the Twitter user timeline API only return up to 3,200 of a user’s most recent tweets. Therefore, our crawling tool continuously monitors all relevant users’ timelines to collect data beyond the 3,200 limit. Note that our approach cannot go beyond the limit for historical data, but rather is a way to circumvent the limit for future tweets.

e) Generating term frequency and co-occurrence networks

From the collected tweets, we calculated the term frequency statistics of the keyterms that we are interested in at both national and state level. The list of keyterms includes not only gender identification terms but also terms that are relevant to the discussions of transgender issues such as ‘transphoa’ (i.e., a range of antagonistic feelings against trans* people based on the expression of their internal gender identity) and ‘HRT’ (i.e., an abbreviation for hormone replacement therapy often used in discussions of gender affirming medical procedures). Moreover, to provide a fair state-by-state comparison, the term frequency statistics were normalized by the number of total tweets of each state. Comparing the term frequency statistics can suggest regional differences in terminology, which in turn can lead to focused hypotheses for further investigations.

Furthermore, we produced co-occurrence networks of the keyterms hoping to discover semantic proximities and the latent structure among them.\textsuperscript{25-27} We formalize a key term co-occurrence network as an undirected weighted graph, \( G = (V, E) \), where each term is a vertex or node \((v_i)\). If two terms co-occurred (in any order) in the same Twitter message, we drew an edge or link \((e_{ij})\) between the two term nodes \((v_i \text{ and } v_j)\), such that the weight \((w_{ij})\) of the edge is set to the number of co-occurrences in all tweets posted by the users of interest. We constructed two co-occurrence networks for each state—one representing the trans* population and the other for the general public (including trans* people).

To assist in the presentation of the results, we built a number of web-based visualizations (http://bianjiang.github.io/twitter-language-on-transgender/). In particular, we used word clouds to depict the representative keyterms; and built interactive network visualizations using a physically-based force-directed graph layout with the Scalable Vector Graphics (SVG)—a language for building rich graphical content,\textsuperscript{28} and d3—a JavaScript library for manipulating SVG objects.\textsuperscript{29}

Results

We collected over 31 million tweets matching the search queries during a 49-day period from January 17, 2015 to March 6, 2015 inclusive. Out of the collected tweets, about 11 million tweets (36.1%) were in English. We were able to extract location information for 141,400 tweets (1.24% of English tweets from 57,997 unique users), which we retained for further processing. Next, we applied the two developed classifiers. We eliminated the tweets that were deemed irrelevant (5,685 tweets from 1,899 users). From the rest of the data set, 56,098 Twitter users were classified as relevant, of which 1,129 users were classified as self-identifying trans*. In addition to the data we collected using the search API, we crawled more than 154 million tweets from the 56,098 relevant Twitter users’ timelines. Out of the 154 million tweets, 532,682 Twitter messages contain one or more of the keyterms of our interest. These 500k tweets represent the corpus we used for language usage analysis.
Table 2 shows the top ten most frequently used keywords across the US on Twitter by trans* people vs. the general public. We present the data on the percentage scale to make the results comparable between the two population groups. In the table the term ‘trans’ occurs frequently because it is part of other keywords (e.g., ‘trans people’ and ‘trans woman’) that we are interested in. For the same reason, ‘trans’ co-occurred frequently with terms like ‘trans people’ and ‘trans woman’. For the purpose of better presentation, we removed any top ranked co-occurrence pairs that contain the term ‘trans’ in Table 2.

As reported in Table 2, the most frequently used terms are similar between users classified as trans* and the general public on the national level. The Spearman’s rank correlation coefficient of the term frequency lists (i.e., the general public vs. trans* people) yields a value of 0.943 (with a two-sided p-value of $8.38 \times 10^{-47} < .01$ significance level) indicating the two lists are highly correlated. On the national level there is a common vocabulary invoked to discuss gender-related issues online.

<table>
<thead>
<tr>
<th>Rank</th>
<th>General Public Terms</th>
<th>Trans* People Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Term Frequency</strong></td>
<td><strong>Co-occurring</strong></td>
</tr>
<tr>
<td>1</td>
<td>trans (32.05%)</td>
<td>#tgirl, shemale</td>
</tr>
<tr>
<td>2</td>
<td>transgender (19.71%)</td>
<td>#tgirl, sissy</td>
</tr>
<tr>
<td>3</td>
<td>cis (6.81%)</td>
<td>shemale, sissy</td>
</tr>
<tr>
<td>4</td>
<td>shemale (3.78%)</td>
<td>shemale, tranny</td>
</tr>
<tr>
<td>5</td>
<td>gender (3.51%)</td>
<td>gender, transgender</td>
</tr>
<tr>
<td>6</td>
<td>transphobia (3.12%)</td>
<td>#tgirl, tranny</td>
</tr>
<tr>
<td>7</td>
<td>tranny (3.11%)</td>
<td>ladyboy, shemale</td>
</tr>
<tr>
<td>8</td>
<td>trans people (2.78%)</td>
<td>#tgirl, ladyboy</td>
</tr>
<tr>
<td>9</td>
<td>#tgirl (2.15%)</td>
<td>gender, gender binary</td>
</tr>
<tr>
<td>10</td>
<td>trans woman (1.96%)</td>
<td>ladyboy, tranny</td>
</tr>
</tbody>
</table>

Table 2. The top ten terms and co-occurring terms tweeted across the United States by the general public vs. trans* for gender identification and discussions of gender-related issues. (*The number in the parenthesis corresponds to the percentage of tweets that contains the term.*)

However, for the sake of developing gender identity questions on intake forms we want to know whether there are reasons to suspect differences among terms used by trans* people at the regional level. Furthermore, since the same intake form is used for both trans* people and non-trans* people, we also want to be able to compare the terminology used by the trans* community with the general population to minimize non-trans* patients inadvertently indicating a trans* status. For further details on false negatives with respect to transgender identification, we direct the reader to The GenIUSS Group: Gender-Related Measures Overview. To gather data for distributional similarity measures, we performed the same term frequency and co-occurrence analysis for each state. Figure 2 compares the word clouds of the keywords used in Arkansas between trans* people and the general public; while Figure 3 depicts the word clouds of the keywords used by trans* people in Arkansas, Florida, Washington, and Kansas. Consider the term ‘genderqueer’ which appears on all three of the sample intake forms in Table 1. An examination of the word clouds in Figure 3 shows that in Arkansas the general public uses the term ‘genderqueer’ (0.46%) more frequently than the term ‘genderfluid’ (0.23%). In contrast, the trans* population uses the term ‘genderfluid’ (0.57%) more often than the term ‘genderqueer’ (0.28%) in Arkansas. Similarly, as shown in Figure 3, the term ‘genderqueer’ while present in Kansas (0.17%) among trans* people, is used less frequently than ‘genderfluid’ (0.34%) and ‘agender’ (0.43%). In Washington, ‘genderqueer’ (0.77%) and ‘genderfluid’ (0.80%) are used with about the same relative frequency. However, in Florida, we see usage that is the inverse of Arkansas and Kansas; ‘genderqueer’ (0.43%) is used more frequently than ‘genderfluid’ (0.25%). Further, ‘agender’ is used with less relative frequency in Arkansas (0.28%) and Florida (0.33%) than in Kansas (0.43%) and Washington (0.46%).

616
Further, consider the term ‘neutrois’, which describes individuals who feel that they have no gender or are gender neutral. This is an example of a regionally specific term whose meaning cannot be characterized as “not exclusively male or female” (similar to ‘agender’), and as such would not be captured by the options of the sample intake forms surveyed. We found that users classified as trans* used the term ‘neutrois’ in only twelve states: CA, FL, GA, LA, MA, MI, MN, NY, PA, TX, VA, and WA. That is, in addition to states such as CA, MA, VA, and WA that are known for having a large identified LGBT population, ‘neutrois’ appears in the Great Lakes states and some of the Southern states.

**Discussion**

*Generating hypotheses about language preference*

While there are no definitive conclusions that can be drawn from this data alone, the findings suggest the following conjecture: intake forms in the southern United States that use ‘genderfluid’ and ‘agender’ rather than ‘genderqueer’ may have better response rates than forms that use ‘genderqueer’.

Furthermore, in light of the findings of the term ‘neutrois’ in the Great Lake states and select southern states, another conjecture is that trans* populations in these states have a higher incidence of ‘neutrois’ in the free-text ‘please specify’ fields.
Similar conjectures can be generated by comparing the co-occurrence networks at the individual state-level. For example, Figure 4 shows the structures of the two co-occurrence networks around ‘genderqueer’ for trans* in Washington and trans* in Florida. Thicker arcs indicate more frequent co-occurrence.

These co-occurrence networks show regional variations in the co-occurrence of ‘genderqueer’. For example, in Washington ‘genderqueer’ co-occurs with terms ‘trans-boy’ and ‘demi-boy’ that fit the Fenway Health characterization of ‘genderqueer’ (“not exclusively male or female”), but also terms such as ‘agender’ and ‘non-binary’ which are gender identities that do not fit this characterization. On the other hand, in Florida, ‘genderqueer’ does not co-occur with terms favoring one end of a binary spectrum, but it does co-occur with ‘non-binary’. From these observations, we can generate the following conjectures: ‘genderqueer’ denotes a broader set of gender-identities in Washington than in Florida, and in both cases it denotes identities not adequately characterized by Fenway Health’s gloss.

These conjectures, however, stand in need of further testing using formal and controlled methods. One of the trouble spots for our research is that some states have very few relevant tweets collected and few users classified as trans*. Delaware, Montana, and Wyoming each only had one user classified as trans* while South Dakota and Mississippi each has only two. While it is likely that this is because there are relatively few trans* persons in these regions using Twitter to discuss issues related to gender identity, it is also possible that trans* related tweets are not captured in our data set because the language used to discuss these terms are not in our list of keywords. We reviewed the raw data captured to date with the current keyterms to find additional keyterms we have missed, but did not find any in this set. These gaps in data point to the need for tools and methods outside of those discussed in this paper for gathering data and testing hypotheses about variations in transgender identity terms and capturing those that are used by people who are less vocal about their gender identity.

Limitations

Our study suggests that social media data sources such as Twitter can expand the range of what can be easily measured and provide new types of information for mining health-related knowledge. However, in addition to big data challenges, Twitter data has its limitations and may not be reliable for answering certain questions. First, although Twitter has a set of feature-rich APIs and a relatively open policy for scraping, collecting relevant data to answer a specific scientific question is not easy. We collected over 154 million raw tweets in less than two months; however, only a fraction of the data (500,000 tweets) was deemed relevant to our study. Second, we found that even with a list of well-developed search terms, the returned data set had many false positives, which affirms the necessity of building classifiers to further narrow the search results. Nevertheless, the process of building classifiers is a tedious process involving manually annotating a large number of tweets to produce a gold-standard training data; and the accuracies of the classifiers were not perfect. In particular, the recall of the second classifier – finding self-identified trans* people – is low (0.261) indicating that we have missed many true positive cases. This might be the reason that we do not have a large enough corpus for trans* people. More sophisticated features\textsuperscript{30, 31} can be
incorporated into the classifiers to improve the performance. Third, the geographic analysis was coarse-grained, providing only statistics on the state level. Although we attempted to geotag tweets with more fine-grained location information at the city level, the result was not satisfactory due to common conflicts in city names (e.g., Springfield, SC vs. Springfield, MA vs. Springfield, IL). Even though Twitter added the capability to record geocodes (latitude and longitude) and introduced new geographic metadata (‘geo’ and ‘place’), there are very few tweets and user profiles we collected with geocodes available. One possible reason for this phenomenon is Twitter users having to give explicit consent to allow software vendors to record their geocodes. Another possible reason is that geocodes are only available if the tweets are sent from devices that have Global Positioning System (GPS) enabled. More sophisticated geocoding techniques20, 32 may be utilized to provide more accurate and finer grained location information. However, there is no direct way to integrate these techniques into our pipeline.

We note that our study is limited by the user demographics available on social media platforms. The users of social media tend to be younger (e.g., 37% of Twitter users are under 30, while only 10% are 65 or older, as of 201433); and there are power users who exhibit a substantially greater quantity of activity than the average user.34 These characteristics are likely to create sample bias and impose limitations on mining meaningful information that represents a broader population. For instance, Twitter data may not be reliable for mining information about senior citizens.

Finally, we recognize that our methods do not capture data from the trans* people who are less vocal about there gender identities on social media platforms. This is an inherent limitation of social media data sources and affects the coverage of the gender identification terms. However, does not affect our conclusion of the prevalence of differences in using gender identification terms in the public. Thus, we limited our investigation here to people who have made an explicit statement on Twitter about their identity. While the results from Twitter mining do not always yield language that is appropriate in the context of clinical care and research — for example, there is a significant quantity of advertisements for sex work on Twitter and discussions of gender-related slurs — Twitter has the potential to provide a comprehensive snap-shot of the language used by self-identified trans* individuals.

Conclusion

This research shows that mining information on social media platforms such as Twitter can yield valuable insights to guide hypothesis generation in the development of intake questionnaires. While the output of this pilot study is insufficient to guide the development of better intake forms, it can be used to generate hypotheses for further testing. Furthermore, this data set can form the basis of future research in transgender health care. By capturing terms in context, we have generated a data set that will allow us to look at contextually sensitive aspects of the term use such as sentiment analysis in future research. Utilizing social networking resources also produces a data set that will allow us to begin investigating the social influences related to health concerns among transgender persons, which is one of the priority research areas identified by the IOM. Finally, our experiences with mining Twitter data in this study yield a good process in dealing with large textual social media datasets.

Acknowledgement

This work was supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064 and the University of Arkansas for Medical Sciences UL1 TR000039. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

mobile Digital Access to a Web-enhanced Network (mDAWN): Assessing the Feasibility of Mobile Health Tools for Self-Management of Type-2 Diabetes

Kendall Ho, MD1, FRCP, Lana Newton, BAA, MEd(cand)1, Allison Boothe, BA1, Helen Novak-Lauscher, PhD1
1University of British Columbia, Vancouver, BC, Canada

Abstract
The mobile Digital Access to a Web-enhanced Network (mDAWN) program was implemented as an online, mobile self-management system to support patients with type-2 diabetes and their informal caregivers. Patients used wireless physiological sensors, received text messages, and had access to a secure web platform with health resources and semi-facilitated discussion forum. Outcomes were evaluated using (1) pre and post self-reported health behavior measures, (2) physiological outcomes, (3) program cost, and (4) in-depth participant interviews. The group had significantly decreased health distress, HbA1c levels, and systolic blood pressure. Participants largely saw the mDAWN as providing good value for the costs involved and found the program to be empowering in gaining control over their diabetes. mHealth programs have the potential to improve clinical outcomes through cost effective patient-led care for chronic illness. Further evaluation needs to examine integration of similar mHealth programs into the patient-physician relationship.

Introduction
Diabetes mellitus is a chronic illness which affects over 2.7 million Canadians1. 90% of those affected have Type-2 diabetes, a metabolic disorder characterized by the body’s inability to produce and/or use insulin effectively. In 2010, diabetes cost the Canadian health systems $11.7 billion. This number is expected to rise to $16 billion by 20201. The majority of these costs are attributed to health care costs (including those for complications such as cardiovascular disease, limb amputations and kidney disease), disability and work loss2, some of which could be delayed or avoided through improved diabetes management1

Though management can take a number of forms, a growing body of evidence suggests that technology-based outreach increases patient health literacy and access to quality care and there is specific evidence to demonstrate that self-management and monitoring can improve clinical outcomes as well as quality of life in patients with diabetes3-5. The extension from generalized technology to mobile technology is logical as almost 50% of Canadians are using mobile devices to access the internet, and an increasing amount of time is being spent on these devices as compared to traditional technologies such as computers or laptops6. Mobile technologies offer a number of benefits including portability and improved accessibility, and are a particularly interesting platform for health related uses due to their low cost relative to other health technologies and due to the availability of health related applications.

Mobile Health (mHealth) research has begun to document evidence of its potential for chronic disease and diabetes management7,8 and current literature demonstrates positive findings with respect to the use of mHealth tools in a variety of health contexts9-12 however a recent environmental scan of fully implemented mHealth programs in Canada that use remote tracking and analysis of clinical data for patients’ management of diabetes identified a scarcity of fully implemented or sustained programs. We therefore conducted a feasibility study, mobile Digital Access to a Web-enhanced Network (mDAWN), to examine the use of mobile technologies (wireless monitoring sensors, social media, and text messaging) in an attempt to address this gap and to better understand the potential of mHealth programs to support both individuals with type-2 diabetes and their caregivers.

Methods
mDAWN was conducted to determine the feasibility and potential benefits of using mHealth technologies to support patients with diabetes patients and their caregivers. The program included three different mHealth “tools”; SMS, an online community, as well as wireless and manual monitoring devices and participants were enrolled for a 3 month period. mDAWN used an iterative approach to examine the use of mHealth technologies from a patient/caregiver perspective, conducting a pilot study (December 1, 2013 to February 28, 2014) to enrolled a group of patients and caregivers to trial the system and refine our protocol and monitoring strategy and a full intervention with program evaluation from May 1st, 2014 to July 31st, 2014.
Participants were recruited from the general public using convenience sampling. The pilot study engaged 24 participants: 13 patients and 11 caregivers (family members or friends identified as health supporters). During the study, two dyads dropped out due to injury and a lack of interest due to technology delivery delays. The second study group was comprised of 43 participants, 26 patients and 17 caregivers.

In order to be eligible to participate in the mDAWN program, participants needed to be able to receive text messages and access the internet. They also needed to speak English and be at least 18 years of age. Patient participants were offered an honorarium of $100 and caregivers were offered an honorarium of $75. Initially, it was required that patients select a caregiver to participate with them in the study; however, this requirement was removed as potential participants reported that they did not necessarily view the presence of a caregiver as important to their type-2 diabetes management. All participants completed an informed consent process.

Participants took part in a 3-month study during which they received access to a website with health resources and a semi-facilitated discussion forum, bi-weekly text messages with health tips and challenges, and three monitoring devices including a wireless blood pressure monitor which measured systolic and diastolic blood pressure as well as heart rate, a wireless weight scale, and a manual blood glucose meter.

The blood pressure monitor and weight scale were purchased from blipcare (www.blipcare.com), a subsidiary of the Carematix company, as they fulfilled the study target of consumer-marketed technology with wireless connectivity capabilities. Measurements from these devices were automatically uploaded to a secure website, accessible to patients and their caregivers, when taken and could be viewed in a number of different graphical and report formats. The blood glucose meter was standard and patients were asked to measure their blood glucose twice daily and enter these results into an electronic tracking form designed by the study team. Patients and caregivers also had access to a web-portal with health resources and a semi-facilitated discussion board. Access to this webportal was limited to participants and study team members and discussion were started both by study team members and by participants. In addition to this, participants received text messages twice a week over the three month intervention period; the first message was designed to be informative, including a health tip and a link to a matching resource on the website and the following text provided a challenge or a question on the same topic, encouraging participants to interact with other study participants using the web-portal.

During the three month study period, evidence of the mDAWN’s impact was gathered from four sources:
1. Self-reported health measures
2. Physiological outcomes in weight, blood pressure, heart rate, and blood glucose
3. Program Cost and Feasibility
4. In-depth post-study interviews with participants

**Self-reported health measures**

Patients and caregivers were administered an online survey at the beginning and again at the end of the three month interventions. Nine different measures, selected by a literature scan of similar programs and through review by the clinicians on the research team, were administered to participants. We chose a large number of measures to scrutinize which ones could best reflect the impact of mHealth activities. A statistical and theoretical analysis led to the selection of the following measures for pre-post comparison:

Health Distress: measures how much distress a person experiences directly related to their health in the past month. It is more specific to distress stemming from health issues rather than a general measure of anxiety or depression. The lower the score, the less distressed a person is about their health.

eHealth Literacy (eHEALS): measures the ability to read, understand and communicate about health information in regards to internet resources to make informed health decisions. High eHealth literacy is indicated by a high score on this measure.

Diabetes Empowerment – Short Form (DES-SF): measures psychosocial self-efficacy of people with diabetes, dissatisfaction, readiness to change, and goal attainment. The higher score, the better the patients’ sense of empowerment when it comes to managing their diabetes.
Quality Adjusted Life Year (QALY) Index: a measure of disease burden that assesses the quality and quantity of life lived in one year ranging from 0 to 1.0, with 1.0 being a full quality life year lived. It is typically used in assessing the monetary value of a medical intervention. The Euroqol 5D-5L (EQ-5D-5L) and United Kingdom look-up table was used to obtain the QALY values for this study.

Two measures were not included beyond a preliminary analysis, the Patient Activation Measure (PAM) and Patient Health Questionnaire (PHQ-9) due to high correlations which indicated overlap with other measures. The PAM is a broad measure of confidence with health management and is not diabetes-specific. The PHQ-9 measures depression, but not in relation to a person’s current view of their health otherwise. Social support scales were excluded as patients did not identify with being in need of a caregiver. (Measures of social support that were considered: Multi-dimensional Scale of Perceived Social Support (MSPSS), Zarit Burden Interview, and Kingston Caregiver Scale).

Physiological outcomes in weight, blood pressure, heart rate, and blood glucose
Data was analyzed from patients’ weight, systolic blood pressure, and blood glucose measurements. Additionally, patients visiting their physicians before and after the study to provide a measure of HbA1c. Weight and systolic blood pressure data were automatically uploaded to the Carematix website when a patient took a measurement. The Carematix website had a patient facing site where the participants could securely sign-on and review their self-monitoring data as well as an administrative side where the research team could view data from all participants.

Participants were asked to obtain their glycated hemoglobin (HbA1c) from their physician prior to the mDAWN study and at the end of the three month program. Blood glucose measurements were collected by patients manually and then emailed to the research team. Patients were instructed to measure their blood glucose twice a day at alternating times every other day. Specifically, patients would measure when they woke up in the morning and before they went to bed on one day and then the following day they would measure before lunch and then again before dinner. This resulted in an aggregate view of blood glucose at four times a day: (1) at wake-up (morning), (2) before lunch time (midday), (3) before dinner (evening), (4) before bed (late evening).

Program Cost and Feasibility
An environmental scan was conducted to identify research or clinical programs similar to mDAWN for comparison in implementation and costs. The search was limited to programs in Canada that use technology to improve health outcomes for people with diabetes by enabling remote tracking and analysis of clinical data.

This scan focused on three information areas: peer-reviewed literature (from 2008 to 2014), grey literature and key informant outreach. Grey literature included a review of national and provincial health organization publications, press releases, internal reports and other online documentation. Sixty inquiries were made of key informants around organizational use of mHealth programs for diabetes management.

Patient participants were also given a cost-feasibility survey at the end of their post survey that asked about annual household income, what devices the participants’ were most likely to pay for, how much they would pay to use mHealth devices to manage their type-2 diabetes and the perceived value of an mDAWN-like program.

In-depth post-study interviews with participants
Patient and caregiver participants were invited to take part in a structured interview about their overall program experience with mHealth and how it could fit into their interactions with the health care system as well as specific experiences with the program content (resources, text messages, discussion board, and monitoring devices).

Results

Self-Reported Health Measures
Two pre and post-test comparisons were done, one for patients only, and one for patients / caregiver dyads. Patients’ pre and post test scores were compared on the selected self-report health measures, HbA1c, health distress, diabetes empowerment (DES-SF), and eHealth literacy (eHEALS). Dyads were analyzed for comparisons between patients and caregivers as well as pre and post-test measures on health distress, eHealth literacy (eHEALS) and QALY.

A one-way within groups multivariate analysis of variance (MANOVA) was conducted to assess patient pre and post - test differences on HbA1c, health distress, diabetes empowerment, and eHealth literacy. There was a
statistically significant difference between pre and post testing, Wilks’ Lambda = .391, F (4, 11) = 4.28, p = .025. The effect size for the overall model using partial eta squared was .61 indicating a very large effect. Each dependent variable in the model had a partial eta squared of over .138, which is a generally accepted criteria (Cohen, 1988) for a large effect. HbA1c, health distress and diabetes empowerment scores all saw changes in a positive direction, with HbA1c and health distress lowering and diabetes empowerment increasing. HbA1c decrease is not only statistically significant, but also clinically significant in dropping from 7.41 to 6.77 (see Appendix A, table 1), with 7 as the clinically acceptable upper limit of normal. eHealth literacy decreased from pre to post test. See Table 1 in Appendix A for the means, standard deviations, and partial eta squared of each variable.

A mixed within and between groups MANOVA was conducted to assess differences between caregivers and patients on pre and post-tests of health distress, eHealth literacy, and quality-adjusted life year (QALY). There was a statistically significant interaction between patients and caregivers across pre and post-tests, Wilks’ Lambda = .533, F (3, 12) = 3.5, p < .05. The partial eta squared was .467 indicating a large effect. Patients reported a significant decrease in health distress, whereas caregivers reported an increase with a large effect size of .405 (partial eta squared). Similarly, QALY scores increased for patients and decreased for caregivers with a moderate effect size of .079. Both patient and caregiver eHealth literacy scores decreased, but with no effect size to report (partial eta squared = 0). See Table 2 in Appendix A for patient and caregiver means and standard deviations.

Physiological Outcomes
Weight and systolic blood pressure were each analyzed with a one-way repeated measures analysis of variance (ANOVA) and blood glucose was analyzed using a two-way repeated measures ANOVA to examine measurement time and changes across the three months of the intervention.

There was no significant decrease in the group’s average weight (Wilk's Lambda = .859, F (2, 14) = 1.15, p = .346), though the group lost an average of 3.51 pounds. Additionally there was a significant negative correlation between number of weigh-ins a participant completed and their weight difference by the end of the study (r = -.69, p < .01) indicating that the more a participant weighed themselves, the more weight they could be expected to lose.

There was a significant decline in systolic blood pressure across the three months, Wilk's Lambda = .49, F (2, 11) = 5.7, p < .02. A multivariate partial eta squared .51 was found, indicating a large effect. Further examination using a trend analysis showed that the significant drop occurred from month 1 to month 2 with systolic readings climbing back up again slightly in the third month (quadratic trend F (1, 12) = 7.79, p < .016, partial eta squared = .394).

Blood glucose data was analyzed according to the four measurement times (morning, midday, evening, and late evening) across the three mDAWN months to elucidate if there was a difference over the three months in each of the four measurements. Data was analyzed using a two-way repeated measures ANOVA. There was no interaction effect for time and month (Wilk’s Lambda = .80 F (6, 60) = 1.15, p = .35).

Program Cost and Feasibility
The environmental scan of diabetes-focused mHealth programs revealed that there are very few published evaluations of remote monitoring programs- while a number of pilot programs featured remote monitoring, none appeared to have a mobile component or reported outcomes beyond the pilot stage. Outreach to key informants supported this finding, with most stating that they were unaware of any mobile health programs for diabetes being used within their organizations. Many were familiar with mobile apps for diabetes management, and multiple respondents expressed interest in learning more about the outcomes of mDAWN, however, no baseline was found to provide a cost comparison point for the mDAWN program.

Patient participants in the second study group also completed a cost survey at the end of the program. In this survey, participants were provided with information about the variable costs of the program and asked questions designed to help the study team understand the perceived value of a program such as mDAWN. 24 participants answered the cost survey, though not all respondents answered all questions. Key findings are illustrated in Table 4.
Household Income

<table>
<thead>
<tr>
<th>Income Range</th>
<th># of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to $20,000</td>
<td>1</td>
</tr>
<tr>
<td>$20,000 to $40,000</td>
<td>4</td>
</tr>
<tr>
<td>$40,000 to $60,000</td>
<td>7</td>
</tr>
<tr>
<td>$60,000 to $80,000</td>
<td>3</td>
</tr>
<tr>
<td>More than $80,000</td>
<td>6</td>
</tr>
</tbody>
</table>

The total cost per person of the mDAWN study is $25 per month for the monitoring service, $35 a month for blood glucose test strips, and $365 for the monitoring devices (weight scale, blood pressure monitor, and blood glucose meter). If you were to purchase an mDAWN-like system yourself, you would spend $365 at start-up and then $60 a month after that. Please choose the statement below that best describes how the program cost suits you.

<table>
<thead>
<tr>
<th>Statement</th>
<th># of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am willing to pay for this, but I can't afford it.</td>
<td>6</td>
</tr>
<tr>
<td>I can't afford it and I wouldn't want to pay for it if I could.</td>
<td>4</td>
</tr>
<tr>
<td>I am willing to pay for this, but it would be a financial burden.</td>
<td>3</td>
</tr>
<tr>
<td>I would only want to pay for parts of this (for example, just the weight scale).</td>
<td>5</td>
</tr>
<tr>
<td>I can afford this and I would pay for it.</td>
<td>4</td>
</tr>
<tr>
<td>I can afford this and it would save me money in other ways. How would it save you money?</td>
<td>0</td>
</tr>
</tbody>
</table>

If the British Columbia Medical Services Plan (MSP) covered some of the cost for home monitoring systems, would you take part and what would be the most you could pay?

<table>
<thead>
<tr>
<th>Option</th>
<th># of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, but the maximum I could pay is $0</td>
<td>10</td>
</tr>
<tr>
<td>Yes, but the maximum I could pay is $40 - $60 a month</td>
<td>6</td>
</tr>
<tr>
<td>Yes, but the maximum I could pay is $60-$80 a month</td>
<td>1</td>
</tr>
<tr>
<td>Yes, but the maximum I could pay is $80-$100 a month</td>
<td>2</td>
</tr>
<tr>
<td>Wouldn't participate even if costs were covered.</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 4. Cost Survey Results

**In-Depth Post Study Interviews with Participants**

All participants indicated an overall positive experience with the mDAWN program, finding their participation beneficial. Common themes that emerged from the participant interviews were the creation of self-care routines and habits, increased self-awareness, and participant empowerment. Participating in the mDAWN program increased awareness of diabetes and of the self, related to diabetes. Particularly, participants expressed an increased understanding of how their lifestyle can affect their glucose control, weight or blood pressure measurements. More importantly, through their experiences in the program participants were able to connect and apply the information provided to their own lives. In turn, making these connections allowed participants to effect positive changes in their lifestyle including exercise and eating habits and to achieve goals such as reducing HbA1c levels, weight or blood pressure measurements.

Participants’ experience in the program also increased their sense of empowerment. Throughout the program participants gained an appreciation that they were ultimately responsible in caring for their diabetes. Participants felt that the program provided the right amount of accountability, incentive, and support to allow them to achieve their goals. Some participants indicated that participating in the program increased their feelings of control and level of confidence in their ability to care for themselves. Being generally participant-led, the mDAWN program gave participants the tools and knowledge while still allowing considerable freedom. This allowed participants to take ownership over the positive changes they achieved while participating in the program.
This new sense of self-confidence carried over into their expectations of future interactions with their doctor, as patients noted that they were now able have a more knowledgeable conversation and were more confident in knowing which questions to ask. In essence, participants felt they were now more “activated” patients who would be more involved in making decisions about their care with their doctors.

Discussion

The mDAWN patient group was able to achieve lowered health distress, HbA1c, and systolic blood pressure as well as an increase in diabetes empowerment over the course of the 3 month study. During interviews, the group reported that they were able to achieve their health goals through a bolstered sense of empowerment. Additional findings in our study based on pre- and post-study self-report health measures are worthy of further consideration: the decrease in both patients’ and caregivers’ perceived eHealth literacy and the rise in caregiver distress. The study team can hypothesize that the decrease in eHealth literacy may be a case of participants ‘not knowing what they don’t know’ at the beginning of the study - feeling that they had a good handle on how technologies have then discovered new digital ways to improve health through the mDAWN program. This “broadening of horizons” could have affected their perception of their own eHealth literacy. The rise in caregiver stress found in the study might be caused by the more constant linkage of them with their patients with potential 24/7 connectivity, thereby triggering a sense of rising responsibility. Recent studies and news reports indicate that caregiver burden can be a source of negative health outcomes, and our current findings may suggest a more nuanced view: that taking on or imposing the role of caregiver may be a factor in negatively impacting health for the caregivers themselves. Further research should be conducted in order better understand these results.

In terms of cost, participants largely identified the mDAWN program as providing good value for money and indicated that hypothetically, they would be willing to pay the fees associated with a program like this. While willingness exists, more than half of the participants (56%) also indicated that based on their current income they would be either unable to afford the program, or that they would find it to be a significant financial burden. Questions regarding cost sharing between the patients and the health system funder, in our case the provincial medical program, found that 52% of participants interested in continuing the program would only find it feasible to do so if the cost was zero. It should further be noted here that over half of mDAWN participants had a household income of less than $60,000/year (below that of the national average). This suggests that in order for a program like this to be successful and feasible for all participants, cost-sharing with the health-system or subsidies must be considered. Cost considerations in future studies should include considering how to build a cost model comparison point which is flexible to changes in technology development that impact accessibility and price, as well as exploring strategies which support sustainable and affordable programs.

Results from the mDAWN feasibility study suggest that the mDAWN system, in the context of type-2 diabetes self-management, can improve clinical outcomes and overall wellbeing. Key aspects of the program’s success include the synergistic use of physiologic sensors and social media. This combination of monitoring devices and a secure social media platform empowered participants by providing them with a unique combination of autonomy and connectedness. The monitoring devices allowed patients to quantify their own health data and apply it recognize patterns in their health and understand how lifestyle choices were affecting their wellbeing. Sharing insights on the social media platform provided opportunities to learn from others, share and celebrate success, and created a sense of accountability to a community. This unique combination of autonomy and connection to a peer group was highly valued by participants and identified as a key part of increasing patient empowerment.

This finding supports the validity of the behavioral change model put forward by Bandura (1986) which asserts that that three key sequential steps are needed to affect behavior change, self-monitoring, self-evaluation, and behavior modification, and that these steps work in synchrony and in a continuous cycle of positive reinforcement. Our findings in this study may be applicable to not only diabetes self-management, but also chronic disease self-management in general.

Limitations

The mDAWN program explored the feasibility of a mHealth based program for type-2 diabetes. A convenience sample was used which resulted in a study group with high pre-study levels of motivation, eHealth literacy, and activation and relatively low levels of health distress and perceived burden. While the study has addressed the key
research questions, the findings of this study cannot at this point be generalizable to the general population. Further research needs to be conducted to fully understand:

- What value can a mHealth program provide to patients with low levels of motivation or health engagement? Recruitment for mDAWN attracted participants who were interested in improving their health. While mDAWN supported improved outcomes for these individuals, we are unable to say how this program might work with a less-motivated patient group.
- Can those who might not be as familiar with mobile health technologies still find this system useful for them? For those who use social media relatively less than the mDAWN study group, is the social media component of this program of value to help them connect with their peers?
- How can health professionals add value to this monitoring system? Our study did not include health professional participation and focused on the value of mhealth tools when used independently by patients. This is in contrast to many conventional remote patient monitoring studies which commonly put patients in a passive role of being managed by health professionals. While both strategies have benefit, there is a gap in knowledge regarding both the value that mhealth tools might offer to the health professional-patient partnership model, as well as a lack of information regarding what would be required to fit this sort of system into existing workflows. Further research is necessary to understand how health professionals can best support patients in the use of mHealth tools as part of an integrated self-management plan.

The next step in this evaluation would be to conduct a clinical trial to address the questions above, and also to allow for the quantification of the benefits seen in this feasibility study.
Appendix A

Table 1 – Statistics for patient self-reported health measures MANOVA

<table>
<thead>
<tr>
<th></th>
<th>Pre Mean</th>
<th>Pre SD</th>
<th>Post Mean</th>
<th>Post SD</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health Distress</td>
<td>12.40</td>
<td>4.82</td>
<td>9.53</td>
<td>3.29</td>
<td>.368</td>
</tr>
<tr>
<td>HbA1c</td>
<td>7.41</td>
<td>1.42</td>
<td>6.77</td>
<td>1.05</td>
<td>.279</td>
</tr>
<tr>
<td>Diabetes Empowerment</td>
<td>31.05</td>
<td>2.38</td>
<td>32.5</td>
<td>2.01</td>
<td>.243</td>
</tr>
<tr>
<td>eHealth Literacy</td>
<td>25.04</td>
<td>3.69</td>
<td>23.73</td>
<td>4.24</td>
<td>.221</td>
</tr>
</tbody>
</table>

Table 2 – Statistics for patient and caregiver self-reported health measures MANOVA

<table>
<thead>
<tr>
<th></th>
<th>Pre Mean</th>
<th>Pre SD</th>
<th>Post Mean</th>
<th>Post SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health Distress</td>
<td>12.0</td>
<td>4.85</td>
<td>8.44</td>
<td>3.5</td>
</tr>
<tr>
<td>QALY</td>
<td>.826</td>
<td>.119</td>
<td>.866</td>
<td>.113</td>
</tr>
<tr>
<td>eHealth Literacy</td>
<td>25.89</td>
<td>3.79</td>
<td>23.56</td>
<td>4.59</td>
</tr>
<tr>
<td>Caregivers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health Distress</td>
<td>5.17</td>
<td>2.04</td>
<td>7.0</td>
<td>3.1</td>
</tr>
<tr>
<td>QALY</td>
<td>.802</td>
<td>.046</td>
<td>.787</td>
<td>.064</td>
</tr>
<tr>
<td>eHealth Literacy</td>
<td>30.0</td>
<td>6.13</td>
<td>28.0</td>
<td>3.03</td>
</tr>
</tbody>
</table>
References:


Uncertainty, Case Complexity and the Content of Verbal Handoffs at the Emergency Department

Jan Horsky, PhD,1,2 Edward H. Suh, MD,3,4 Osman Sayan, MD,3,4 Vimla Patel, PhD5

1Brigham and Women’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3New York-Presbyterian Hospital, NY, 4Columbia University Medical Center, NY; 5New York Academy of Medicine, NY

Abstract
Handoffs are known to increase the risk of medical error and adverse events. Few electronic tools can support this process effectively, however. Our objective was to describe the relationship between clinical complexity, diagnostic uncertainty, fit with illness script and the content of case presentations by physicians. We observed the handoff of care for 150 patients during eleven shift changes at a large urban emergency department (ED). Results indicate that as uncertainty about diagnosis and perceived illness script increased, more descriptive detail was conveyed to the incoming physicians. Physicians were concerned primarily with creating a shared mental model of a patient’s clinical state and with describing the expected path to disposition rather than simply passing on data and findings. Electronic tools for ED handoffs should allow adjustment of structure and content to capture complexity and uncertainty appropriately without requiring extra effort for more routine cases that better fit to more standard narratives.

Introduction
Physicians transferring the care of patients to another team at the end of a shift have the challenging task of collecting sets of clinical data with laboratory results, test findings, significant events and gathering other sometimes complex evidence that collectively characterize a patient’s course from complaints, problems and clinical decisions to disposition plans. All information needs to be presented to incoming clinicians as a succinct narrative of interconnected events that can be accurately interpreted in order to allow them to continue care without delay. Research evidence suggests that handoffs engender misinterpretations and omissions that may lead to medical errors and adverse events1,2 but that these discussions also afford the opportunity to consider alternative viewpoints, correct mistakes and propose new tests or interventions.3 Efforts to analyze and improve the process generally assume that its primary goal is a unidirectional and complete transfer of information between individuals and teams on adjoining shifts. However, it is also an opportunity for a dialog in which participants co-construct and negotiate a shared understanding of existing clinical issues and check the accuracy of medical reasoning and prioritize next steps.4

Developing a shared mental model of illness and care by the active involvement of both parties during a handoff is particularly relevant to emergency medicine where adequate and reliable information may not be available and where uncertainty about symptoms, presentation and several possible diagnoses create ambiguities that allow different interpretations.2,5 Interventions and strategies intended to make handoffs more robust and resilient toward misinterpretation and loss of important facts usually emphasize training, standardization and the use of checklists and structured forms as best practices.6 This approach has clearly reduced adverse events in some settings7 and has appropriate application in specific areas of care.

Intrinsic to emergency care, however, are frequently changing priorities, time limitations, the need to treat multiple patients concurrently and large variations in problem complexity and acuity that are encountered and treated on a single shift. When the incoming team takes over, patients are at very different points in their care trajectory. Uniform and inflexible handoff structure is therefore not likely to provide support to clinicians effectively and may even introduce costly unintended consequences.8 For example, in an environment where time and attention needs to be allocated to many patients simultaneously, detailed, exhaustive documentation of routine problems or standard treatments may in effect reduce the opportunity to discuss complex and uncertain conditions or unfolding events.

Theories of cognition distinguish between reasoning strategies that people use in ambiguous, complex and arbitrary situations to develop insight and understanding and those used in predictable or familiar conditions. Problems that can be articulated within clearly defined categories, relationships and hierarchical structures, such as a relatively uncomplicated or routine illness and care history, are generally best evaluated and resolved by the use of paradigmatic reasoning. Standardized handoff tools and templates will likely improve human performance and reduce the occurrence of errors for these cases. By contrast, a narrative mode of reasoning becomes more effective for complex
problems as it organizes knowledge into temporal plots, linking specific events into causal chains and emphasizes consequential connections among them.\textsuperscript{9} Highly structured, static forms of support will not align well with uncertain diagnoses and indeterminate illness scripts. Modern handoff tools will need to support both cognitive strategies equally well.

Current methods of coordinating handoff procedures and organizing communication provide an important foundation for tool development but may be conceptually limiting.\textsuperscript{10} Market innovation has not yet adequately addressed existing design challenges and electronic handoff modules often resemble or mimic paper forms,\textsuperscript{11} providing little meaningful support for diverse cognitive tasks. Studying handoffs from the perspective of uncertainty, complexity and perceived fit to illness script is in our opinion a productive approach to identifying and describing cognitive strategies that are used in different clinical situations and that may need different kinds of support. For example, new insights may inform the design of electronic tools that will be appropriate for the volume and character of information discussed between outgoing and incoming teams for patients with high and low levels of diagnostic certainty or who are at a different stage in their emergency department (ED) trajectory. The goal is to give clinicians the right information in a form that will allow them to make timely and effective interventions when needed and to sustain the care process uninterrupted across shift and unit boundaries.

Our objective was to observe discussions of physicians during handoffs and to analyze how case complexity, uncertainty and fit to illness script affect the content of their patient case presentations. This predominantly descriptive and exploratory study was intended to report findings from a research perspective not frequently present in publications and also to test our assumptions about the relationship between uncertainty levels, clinical complexity and different needs of cognitive support.

**Background**

There seems to be little convergence in clinical research literature on the characterization of a good handoff\textsuperscript{12} and little published evidence on what constitutes best practices.\textsuperscript{13} Many quality measures have been proposed although there is a general lack of consensus on the primary purpose of handoffs and how to best improve the process.\textsuperscript{14} For example, one rating scale assessed information transfer, shared understanding, working atmosphere, and overall quality\textsuperscript{15} and another rated, on a nine-point scale, performance in interviewing, physical examination, humanistic qualities and professionalism, clinical judgment counseling, and organization and efficiency.\textsuperscript{16} Recognizing that there are multiple purposes for handoffs that need to be addressed simultaneously is a critical precursor to quality improvement.\textsuperscript{17} Studies of uncertainty in clinical work also give a somewhat fragmented and incomplete account of its effects and important insights have not been translated to practice.\textsuperscript{18} One taxonomy conceptualizes uncertainty as a multi-dimensional phenomenon with theoretically distinct domains and constructs that are potentially measurable and related to different outcomes, mechanisms of action, and management strategies.\textsuperscript{19} A measure developed to study clinical reasoning strategies during patient visits and derived in part from a cognitive engineering framework includes assessment of uncertainty that refers to how well limitations of available information are recognized, explained and solutions are planned to adjust to the current situation.\textsuperscript{20} Our current framework uses uncertainty for this study in a similar way to analyze its relationship to reasoning and the amount of detail conveyed in narrative accounts of care during handoffs.

Cognitive and human-computer interaction studies in healthcare have shown that team and system interaction,\textsuperscript{21} staff workload,\textsuperscript{22} clinical workflow and the effects of technology on decision making and cognition\textsuperscript{23} need to be investigated if new information technology is to be meaningfully integrated into the process of care. A design paradigm that embodies this comprehensive approach and is employed routinely for the design of safety-critical systems (although still somewhat sporadically for health information technology) is User-Centered Design.\textsuperscript{24} The most common method for gaining insights and understanding of tasks and the environment is direct observation of work and ethnographic studies. Our investigation contributes to the body of knowledge about the handoff process in emergency care and can be used as a starting point in the design of new technology to support clinicians.

**Methods**

We observed clinicians during rounds in the ED and recorded their conversations concerning patient care handoffs. Eleven scheduled shift changes were studied on ten days over a four-month period. On each occasion, a group of two to three outgoing ED residents presented between seven and twenty-two patient cases to an incoming attending physician. The aggregated time of all recordings was 190 minutes (17 minutes on average) and contained conversations about 168 patients. Recordings were transcribed, de-identified (when names were audible) and reviewed by two physicians to correct errors and misinterpretations. Eighteen patients were excluded either because
they had not yet been formally seen or because of poor recording quality. The remaining 150 cases were analyzed in this study.

Setting

The study was conducted in the main adult ED of a large academic hospital in an urban area. The annual census is approximately 88,000 in three clinical districts and a separate fast-track area. Two of the districts have shift changes for all providers twice a day, at 8am and 8pm. The attending physicians and senior residents in the third district have three shifts a day, ending at 8am, 4pm, and at midnight. Emergency medicine residents, those rotating from different services and attending physicians will “round” as a group on all patients in the district. These rounds constitute the usual handoff process between the teams. At the time of data collection, no program for improvement of handoffs had been implemented. Rounds were primarily recorded at 8am and 4pm when the census tends to be significantly lower than when shift change occurs at 8pm and midnight. A fully integrated EHR system was used for virtually all clinical activity.

Table 1 Definition of assessment measures and scale levels

<table>
<thead>
<tr>
<th>Measure</th>
<th>Scale and definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apparent diagnostic uncertainty</td>
<td>0 Unable to determine.</td>
</tr>
<tr>
<td></td>
<td>1 Specific presumptive diagnosis (e.g., “appendicitis on CT, UTI, pneumonia”).</td>
</tr>
<tr>
<td></td>
<td>2 Diagnosis fits into a narrow differential (e.g., sepsis, vaginal bleeding). If unclear, process is well defined for a workup (e.g., chest pain).</td>
</tr>
<tr>
<td></td>
<td>3 Diagnosis appears to be unclear.</td>
</tr>
<tr>
<td>Apparent fit with illness script</td>
<td>0 Unable to determine.</td>
</tr>
<tr>
<td></td>
<td>1 Typical presentation, fits well illness script.</td>
</tr>
<tr>
<td></td>
<td>2 Atypical presentation, but seems a variant of a typical problem.</td>
</tr>
<tr>
<td></td>
<td>3 Does not easily fit into a pattern that is readily recognized.</td>
</tr>
<tr>
<td>Apparent clinical complexity</td>
<td>0 Unable to determine.</td>
</tr>
<tr>
<td></td>
<td>1 Simple problem with few or no complicating features (e.g., isolated laceration).</td>
</tr>
<tr>
<td></td>
<td>2 Moderately complicated problem but not complex (e.g., pneumonia)</td>
</tr>
<tr>
<td></td>
<td>3 Complex elements or a highly complicated problem (e.g., undifferentiated shock).</td>
</tr>
</tbody>
</table>

Analytical constructs and coding scale development

We operationalized the assessment of uncertainty along three axes with 3- and 2-point scales. The constructs were derived from established concepts in medical education and clinical care that are well described in literature and have generally accepted meanings. They were developed and refined iteratively in a series of meetings and conversations. For example, scales were initially proposed to have 5-points but were subsequently simplified so that they were more appropriate for describing variations encountered in the analyzed sample.

Apparent Diagnostic Uncertainty estimates the amount of confidence physicians seemed to have in their understanding of the disease process that lead to the acute presentation. The concept may be defined in ways that are specific to clinical context. In our interpretation, uncertainty about a diagnosis may originate from the lack of expert knowledge, level of training, experience with a particular disease and its presentation, and also from information conveyed by patients whose clinical history may contain ambiguity.

Apparent Fit with Illness Script to typicality of the presentation for a particular complaint or diagnosis is based on a theoretical framework describing how experts prioritize syndrome recognition through comparing and contrasting key clinical features in making a diagnosis. Expert clinicians use illness scripts most of the time in their clinical reasoning since it involves a highly efficient knowledge-driven model of pattern recognition.

Clinical Complexity was an assessment regarding the number of complicated elements and the overall extent of problem complexity classified on a 2-point scale. The complex patient has been described as “one for whom clinical decision-making and required care processes are not routine or standard.” Our characterization of the concept is...
somewhat qualitative as the definition is difficult to translate into quantitative terms. Definitions of each scale level are in Table 1. The acuity of each case and disposition status were also assessed as a secondary descriptive measure.

The amount of descriptive detail in communications about each patient as presented by the outgoing physician was classified into three categories: History of Present Illness and Physical Examination, Diagnostic Findings and Course of Treatment. The categories correspond to standard components of evaluation practice in the ED. They were aggregated into the Informational Content measure and scored on a three-point scale (Table 2). Assessment and classification definitions were developed by two physicians (authors) based on their clinical expertise and detailed knowledge of emergency care and further refined (originally from a 5-point scale) through consensus in a series of meetings and conversations. Audio recordings and their transcripts were scored by an ED physician (author) and results reviewed and revised in collaboration with another physician (author) to assure appropriate coding and classification.

Table 2 Definition of informational content measures and scale levels

<table>
<thead>
<tr>
<th>Informational Content measure</th>
<th>Scale and definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of present illness / Physical examination</td>
<td>0 None present.</td>
</tr>
<tr>
<td>1 Minimal or brief (e.g., chief complaint or a brief history of symptoms only).</td>
<td></td>
</tr>
<tr>
<td>2 Moderate (e.g., linear history or physical findings, recent history, some comorbid conditions).</td>
<td></td>
</tr>
<tr>
<td>3 Detailed (e.g., specific history and physical findings, description extends beyond the present illness).</td>
<td></td>
</tr>
<tr>
<td>Diagnostic findings</td>
<td>0 None present.</td>
</tr>
<tr>
<td>1 Minimal or brief (e.g., only qualitative descriptions such as “labs were ok”, “appendicitis on CT”).</td>
<td></td>
</tr>
<tr>
<td>2 Moderate (e.g., quantitative descriptions of selected tests).</td>
<td></td>
</tr>
<tr>
<td>3 Detailed (e.g., quantitative descriptions of several tests).</td>
<td></td>
</tr>
<tr>
<td>Detail of ED course</td>
<td>0 None communicated, only disposition</td>
</tr>
<tr>
<td>1 Minimal or brief (e.g., status before disposition “couldn't endorse,” definitive treatment given related to disposition or diagnosis, “Lasix for CHF”).</td>
<td></td>
</tr>
<tr>
<td>2 Moderate (e.g., highlights events and treatments influencing current status).</td>
<td></td>
</tr>
<tr>
<td>3 Detailed (e.g., provides thorough detail about ED events and treatments).</td>
<td></td>
</tr>
</tbody>
</table>

The dataset was uploaded to SAS 9.3 statistical software111 and tests were performed on the Confidence and Informational Content sets of variables. Non-parametric analysis of variance for ordered values (Jonckheere–Terpstra Trend Test) was used as a measure of difference between groups with three levels of a predictor variable. The levels of Clinical complexity were analyzed using the Wilcoxon Score (Rank Sums) test.

Results

The assessed acuity of most cases (67%) was found to be of small to moderate likelihood of morbidity and mortality. Nine cases (6%) were considered more serious and for twenty-nine (19%) acuity was not possible to determine. Disposition status at the point of a handoff was to admit (including boarding) for 86 (58%), discharge for 13 (9%) and under evaluation for 51 (33%) patients. Informational detail conveyed by the presenting to the incoming physician increased with rising diagnostic uncertainty and case complexity. Results are provided below. Measures are compared to the History of Present Illness and Physical Exam, Diagnostic Findings and Detail of ED course.
Apparent diagnostic uncertainty

Two patients did not have any history or physical examination discussed and for 25 (17%) diagnostic uncertainty could not be determined from the recordings. The remaining sample of 123 patients was analyzed (Figure 1). Most patients (76, or 62%) either had a narrow differential diagnosis or the process for a workup was well defined. The extent of communicated history and physical exam detail was mostly minimal, with a smaller proportion of patients with a moderate or detailed description. Similar pattern was observed for patients with a presumptive diagnosis. This trend was reversed when the diagnosis was uncertain where most patients (42%) had a detailed description. We found the statistical difference between groups with Presumptive, Differential and Uncertain diagnoses for levels of informational detail about the history of present illness and physical examination to be significant (p=0.0009).

The quantity of Diagnostic Findings in the presentations was mostly minimal for all groups (Figure 2). Equal proportions of cases had no or moderate levels of descriptive detail, and patients with the most detailed explanations represented the smallest proportion in each group. This trend was most pronounced for differential diagnoses that were described only in minimal descriptive detail, for 43 patients (35%). Patients with uncertain diagnoses tended to have larger proportions of detailed description than other groups. Difference between groups with Presumptive, Differential and Uncertain diagnoses was statistically significant for levels of detail about diagnostic findings (p=0.024).

The Detail of ED course was described for all groups in a pattern similar to Diagnostic Findings although differences between the levels of detail were not statistically significant. The sample consisted of 125 patients: 25 (17%) were excluded because the ED course could not be determined from the recordings.
Two patients did not have any history or physical exam discussed and for 44 patients (29%) the fit to illness script could not be determined from the recordings. The remaining sample of 106 patients was analyzed (Figure 3). Most patients had a typical presentation of an illness (61, or 58% of total) and were most likely to have a minimal description of history and physical (29 or 48% of their group). Those with a variant or atypical presentation were most likely to have a moderately detailed description (16, 55% in group) and those with no apparent fit to an illness script a detailed description (8, 50% in group). Statistical difference between groups with a Typical Presentation, Variant and No Fit with illness script was significant for levels of informational detail about the history of present illness and physical exam (p=0.01).

Diagnostic findings were communicated mostly at a minimal level of description and with a higher level of detail for only three patients (2%) (Figure 4). However, patients with no apparent fit of presentation to an illness script tended to have more information conveyed during the handoff. Statistical difference between groups with a Typical Presentation, Variant and No Fit with illness script was significant for levels of informational detail about Diagnostic Findings (p=0.02).

The Course of Treatment in the ED was communicated in a pattern similar to Diagnostic Findings with a slightly larger proportion of detailed descriptions for all groups. Differences in the amount of descriptive detail between groups with different uncertainty levels were not statistically significant, however. The sample consisted of 106 patients as 44 (30%) were excluded because fit with illness script could not be determined from the recordings.
Clinical complexity

Two patients did not have any history or physical exam discussed, for 30 patients (20%) clinical complexity could not be determined from the recordings and five patients with simple problems were excluded as the group was too small for analysis. The remaining sample of 115 patients was analyzed (Figure 5). Clinical complexity was moderate for 53 patients (46%) and 62 (54%) had more complicated problems. Both groups had about the same proportion of minimal and moderately detailed descriptions of history and physical exam but complex patients had more detailed narratives. Observed statistical difference between groups of clinically moderate and complex patients in levels of informational detail about the history of present illness and physical exam approached significance (p=0.058).

Trend patterns and the distribution of descriptive detail in communication was similar for diagnostic findings and course of ED treatment, with higher proportions of detailed descriptions for more complex cases. The difference between groups of clinically moderate and complex patients in levels of informational detail about diagnostic findings was significant (p=0.001) and approached significance for levels of detail about course of ED treatment (p=0.051).

Discussion

This preliminary investigation shows an association between the extent of descriptive content in verbal communication during handoffs and the level of confidence in diagnosis and probable course of further care that the outgoing physician has. As diagnostic uncertainty rises and the fit of the illness to a recognized script is less typical, the amount of detail in the history of present illness, physical examination, diagnostic findings and the course of care increases.

We have observed several trends in the way diagnostic uncertainty changed during a typical ED course. For many patients, there was initially a high degree of uncertainty about the cause of the presentation and about best ways to evaluate before deciding on disposition. As information was gradually gathered a corresponding reduction in the degree of uncertainty usually ensued but could also rise again when new data did not confirm expectations and hypotheses or new complications arose. A scheduled handoff could happen anywhere on this trajectory and clinicians would need to reason about the same patient differently. An innovative handoff tool would allow them to select the right modality to support the creation of their report or guide their discourse. Patients can be signed out from one provider to the next at virtually any point along the course of evaluation and treatment, at points when uncertainty may be at its peak, its lowest level or anywhere in between. Our analysis has demonstrated that there is a tendency towards giving fuller and more complete accounts of information when the degree of uncertainty is higher.

Diagnosing a patient with a high degree of confidence is generally of less concern to ED clinicians than the process of reaching a disposition. In the context of emergency care, a definitive diagnosis is often elusive, and the constant influx of potentially critically ill patients makes time and cognitive resources precious commodities. There is rarely a complicated plan for treatment that needs to be conveyed, as priorities are in addressing immediate, most critical and time-sensitive needs of new patients seeking care. Incomplete information is a routinely expected condition as are rapid changes and new developments and physicians typically view their role as stabilization and disposition rather than securing definitive diagnosis and management. It is therefore important to emphasize in the handoff process the aspect of collaborative discussions and review. Our contention is that ED physicians are not primarily concerned with transmitting sets of data about patients from one to another in handoffs. Rather, they engage in the process of
developing a shared mental model of each patient’s current state and expected path that can be quickly and accurately adopted for care. These mental models function similarly to individual illness scripts, but encapsulate the entire process of evaluation and disposition, including the prospects of multiple possible outcomes.

When the presenting physician feels uncertain about which pre-defined ED “process script” is appropriate for a patient, the handoff cannot consist only of heuristics but is necessarily more detailed. This allows the incoming team to more explicitly contribute to the understanding of the patient’s state and invites their input into the current plan of management. The narrative of the patient encounter becomes much more valuable to the incoming physicians in constructing their own understanding of the situation, especially where there is asymmetry between the roles of those giving and receiving a report, as it exists for example in teaching hospitals. A form of a handoff decision support would help both parties to navigate through this situation and perhaps provide patient-specific advice to guide their discussion. However, when patients do fit well onto an established process script and expected trajectory, the value of adding additional detail to the handoff is questionable. Required data are usually available in the EHR already.

Rounds were conducted by walking from patient to patient, carrying printouts of the ED status board that includes the patient name, age, and chief complaint as documented in triage and boxes that indicate if laboratory tests, radiology studies and nursing orders are completed or pending. However, these indicators were quickly out of date and were considered unreliable. There were also columns that show the current disposition, and if planned for admission, the assigned admitting team. This information is also not reliable and is usually verbally reviewed. Only one ED attending routinely conducted rounds with a mobile computer workstation. Mobile technology such as tablet computers may serve as a convenient technology platform for handoff tolls so that all parties have access to updated, shared information that is unobtrusively present during discussions.

Our emphasis on the changing rather than stable components of a handoff process is not contrary to the current effort to standardize handoffs in order to achieve more successful transfer of information. Consistency and predictability are core design principles that have been known improve communication speed, clarity and to reduce errors of omission. Checklists or other structural constants would be essential in the ED to quickly and accurately document routine cases with low complexity and high levels of certainty in diagnosis and disposition plans. We are proposing that often the ebb and flow of information is a meaningful variation and when the reasoning of clinicians follows a strategy that needs to account for ambiguous, missing or uncertain information, standardization would not adequately fit the model. Designers would need to include this source of variation into their tools to make them more flexible in supporting the sharing of mental models in authentic conditions of emergency care delivery.

We are not aware of any published tools or programs that explicitly address this understanding of the handoff and believe that a valuable direction for future investigation would be continued research into the nature of the types of information conveyed in the handoff and their usefulness. However, it seems apparent that one size does not fit all, as it were, when sign-out tools and processes are concerned.

Limitations

We did not have access to written and electronic information that was used many physicians during rounds and for documentation of care. Verification and comparison of verbal and written communication was therefore not possible which sometimes limited our insight into the clinical cases that were discussed. The study design was somewhat limited by our focus on the verbal report of the outgoing physician and did not capture the understanding of the incoming physician. This would be an important aspect of the mental model sharing description and would need to include interviews with physicians and other forms of data collection.

Observations were limited to one institutions so the findings may not be directly relatable to other emergency departments, particularly in non-teaching hospitals. Our exclusive priority was on physicians and relevant communication between nurses and other clinical staff in the ED that is also critical to the process was not captured. Our estimates of the stage of the ED process care in which a particular patient was at the time of handoff are often not accurate. This knowledge would be important to validate our assumption about variations in uncertainty along the course of care.

Conclusion

Evidence supporting medical decisions is often shared among clinicians in a written form (electronic or otherwise) as a part of diagnostic and care planning documentation. In the time-constrained environment of emergency care, verbal presentations and discussions during rounds and handoffs are essential for developing a more comprehensive understanding of findings and symptoms and for increasing confidence and certainty through consensus, especially in
the absence of complete information or distinct manifestations of known illness trajectories. Both forms of communication are necessary as they seem to have complementary roles in refining and sharing knowledge that will be the basis for further treatment decisions and interventions. Electronic handoff tools that could direct and focus verbal discourse by allowing clinicians to identify and point out missed or possibly misinterpreted evidence from notes or to elicit opinions and clarifications from other parties may be more effective in safeguarding the integrity of communicated information than a simple demand for completeness regardless of case complexity or apparent good understanding of the patient case by those involved. This view is framed in theories of cognition and management of uncertainty under time-constrained conditions and situated in a larger socio-technical context. Better understanding of the process will have positive implications for both training and design of rounding and handoff tools specific for emergency care physicians.

The importance of understanding handoffs is well recognized by the clinical research community. A well-executed handoff, in our view, is not necessarily one with the most details or high levels of completeness. Rather, the ability of the incoming clinician to make well informed and timely interventions when a patient’s state suddenly changes is seen as a high benchmark of safe and quality handoff process. We hope that our work will provide insight and guidance for the development of modern electronic support tools and best practices for effective and safe handoffs of care in the ED.

Acknowledgements

Our thanks are due to the many clinicians who let us observe their work on numerous occasions and were generous with their insights and expert advice, and to our research associates who recorded their conversations. We are also grateful to the James S. McDonnell Foundation for funding this study.

References

Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study

N. Hosseini, PhD¹, M.Y. Sir, PhD¹, C.J. Jankowski, MD², K.S. Pasupathy, PhD¹
¹Health Care Policy & Research, Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic
²Department of Anesthesiology, Mayo Clinic

Abstract
Operating rooms (ORs) are one of the most expensive and profitable resources within a hospital system. OR managers strive to utilize these resources in the best possible manner. Traditionally, surgery durations are estimated using a moving average adjusted by the scheduler (adjusted system prediction or ASP). Other methods based on distributions, regression and data mining have also been proposed. To overcome difficulties with numerous procedure types and lack of sufficient sample size, and avoid distributional assumptions, the main objective is to develop a hybrid method of duration prediction and demonstrate using a case study.

Keywords
Classification, prediction, hybrid method, regression, surgery times

1. Introduction
Accurate prediction of the duration of surgical procedures is necessary to meet the needs of stakeholders. The duration of the procedure is critical in determining optimal schedules, and to reduce delays for patients and providers as well as reducing overtime and under-utilization of operating rooms (ORs) for administration. As surgeries get scheduled, the surgery duration is estimated and an appropriate day is selected to schedule the surgery. Currently most hospitals use software designed by commercial surgical scheduling systems from EMR vendors such as Cerner, Epic, etc. Both traditional method of surgery prediction and those used by commercial software used in most hospitals is based upon a moving average of previous cases, based on surgeon and procedure codes. This warrants the need for a scientific method to predict duration. With the increasing amount of data in healthcare and the need for making improvements in the hospital industry, it is critical to use more efficient methods of surgery prediction that can improve system performance and encourage physicians and hospital staff to not only think about the success of their own practice as an individual surgeon but also think of improving the performance of the hospital as a system. In this study, we consider data from a large hospital and propose a hybrid method for predicting surgery duration times. Our method is consisting of two steps 1) classification and 2) prediction. In prediction portion, we explore two different methods and discuss the performance of each method.

2. Literature Review
To provide a more scientific mechanism for prediction, various approaches have been proposed. While predicting surgical durations, there are two main streams of research. The first stream attempts to find the best fit among known distributions (most commonly normal [1] and lognormal [2]–[6]) and use these fitted distributions in order to characterize variability in surgical procedures and predict durations. In the second approach, researchers build statistical models to predict surgical procedure durations and identify critical factors that influence variability in these durations.

Among those studies relating to distributions, Strum et. al. analyzed a large dataset of clinical cases and concluded that fitting a log-normal model for each Current Procedural Terminology (CPT) code-anesthesia combination provides accurate predictions for procedure duration [7]. In a follow-up study, Strum et. al. showed that the lognormal distribution provides a better fit than normal distributions for modeling procedure durations having exactly two CPT codes [8]. With such an approach, the distribution is used in scheduling instead of a single value. This approach is good because it considers stochasticity by using distribution, however it is clear that there is a lot of variability within procedures of a certain CPT code in terms of complexity and that the single criteria CPT code is not an accurate indicator of surgery prediction. Relying on the lognormality assumption for procedure durations, Dexter, and Ledolter develop a Bayesian method to calculate prediction bounds for procedure durations [9]. Stepaniak et. al. use a three-parameter lognormal model for predicting the procedure durations of CPT-anesthesia
combinations including surgeon effects and show that their model can significantly reduce prediction errors and therefore operating room (OR) inefficiency [10].

On the other hand, with model building, due to the large number of CPT codes, Strum et. al. built a separate five-factor main-effects linear model for each CPT code [11] using logarithm of surgical time (i.e., the time from incision to closure) or the logarithm of total procedure time (i.e., the time from when the patient enters the OR until he/she emerges from anesthesia) as the response variable and surgeon, anesthesia type, American Society of Anesthesiologist (ASA) risk class, patient gender, and patient age as the explanatory variables. Eijkemans et. al. developed a regression-based prediction model with the logarithm of the total OR time, defined as the time from patient entry into the OR room until the patient is moved out of the OR, as the response variable [12]. Besides the surgeon’s estimate, they considered a large set of additional factors, divided into three classes, including operation characteristics (e.g., the number of separate procedures and whether it is a laparoscopic procedure), team characteristics (e.g., number and experience of the surgical team), and patient characteristics (e.g., age, sex, body mass index, previous hospital admissions) and determined their significance when added as a single factor to the base model, which only included the procedure type as a random effect. Motivated by the fact that CPT codes are among the factor with the highest predictive power for procedure durations, Li et. al. developed a general regression-based predictive model with multiple CPT codes as dependent variables [13]. They developed a grouping procedure to identify CPT codes that always appear together in order to construct a full-ranked design matrix for the regression model. Kayis et. al. developed a regression model, which adjusts a commonly used base estimation method using procedure-surgeon specific last five cases, using operational (e.g., order of surgery, OR assignment, surgical staff) and temporal factors (e.g., day, month, time of day) [14]. Due to the larger number of explanatory variables, they used an elastic-net regularized generalized linear model. Their model results in improved mean absolute deviation, especially for cases with long durations.

Some authors investigated use of data mining techniques to predict procedure durations. Combes et. al. proposed a knowledge discovery in databases (KDD) framework. Within the data mining step of this framework, they developed and compared two data mining methodologies, namely rough sets and neural networks, using patient related factors (e.g., administrative data, previous medical history) and surgical environment (e.g., surgeon, type of anesthesia) as explanatory variables [15]. Based on factors related, the patient and surgical environment (including patient age, experience of surgical staff, type of anesthesia) within an ophthalmology department, Devi et. al. developed and compared the performance of three methods: 1) adaptive neuro fuzzy inference systems (ANFIS), 2) artificial neural networks (ANN), and 3) multiple linear regression [16]. Using duration estimates, they solved a mixed-integer programming problem to optimize surgery schedules with an objective of minimizing overall completion time (i.e., make span). Their numerical experiments indicated that ANFIS outperforms other methods. Instead of predicting duration of individual surgical procedures, some authors study the completion time of a series of surgical procedures (also referred to as operating list) in the same operating room on a given day. Dexter et. al. proposed a regression model to predict the completion time of a list with the number of surgeon-procedure combinations with the list as independent variables [17]. Pandit and Carey used a questionnaire of surgical staff (including surgeons, anesthetists, and senior nurses) to estimate the duration of procedures and subsequently applied the average of these estimates to predict the completion time of historical lists. They concluded that even though estimates from surgical staff are accurate in predicting the completion time of operating lists, a substantial number of lists were overbooked [18]. In a related work, Pandit and Tavare developed a method for calculating the probability that a list will finish within its scheduled time [19].

3. Method and Case Study
As previous research suggests there are several factors that impact surgical times. In this study we received data from a large hospital system. The data includes several fields including some general fields and some patient specific information (we discuss the details related to the data in next section). One of the fields that show to be impacting surgery duration is procedure code. Our data shows the record of 2000 procedure codes during the study period. Therefore, the first step in our proposed method is to statistically reduce the number of sub factors for this field. Our proposed method consists of two steps as part of the model and an evaluation to assess the performance of the model (as shown in Figure1). In first step, we use classification to group procedure codes and to reduce the number of sub-factors for the field of procedure code. The next step is prediction; in this step we develop two separate regression models for procedure duration prediction using classical least square linear regression with main factors included (LIN) and stepwise regression (STEP) where main factors and second level interactions are included (we note that the stepwise with more levels of interaction did not add value, therefore were not considered).
We then evaluate the model by comparing prediction results from LIN and STEP against the baseline. The hospital system prediction (baseline) currently uses the moving average of 5 to 10 previous cases of same surgeon for that procedure code. The system then allows the scheduler to adjust this value. Therefore, the recorded value for procedure time in the system is the adjusted system prediction (ASP) and that is the value we use as baseline. The adjustment scheduler makes to the moving average value is reflected by clinical situation of patient (such as preexisting conditions, etc.) and the complexity of the surgery which we do not have any fields available for that in electronic data. This may seem that we set an unfair baseline to compare with. However, we feel confident that if our model can outperform this baseline, then with additional data fields that will be added in future, our model can perform even better.

![Figure 1. Research Method](image)

### 3.1. Data Structure
A total of 63,254 surgical procedures performed by 234 surgeons over 39 months at a large academic health system with a total of 60 ORs were included. The data fields that show to impact surgery times and are included as part of the electronic data records are specialty, priority, ASA class (American Society of Anesthesiologists score – preoperative evaluation of patient physical status), age, encounter class, and procedure code. We also use fields such as actual surgery start and stop times, actual OR start and stop times, and scheduled OR start and stop times for evaluation purposes. Tables 1 and 2 show how the patients are distributed in relation to factors priority class and ASA code. We also note that there are total of 2000 procedure codes, and 30 specialties associated with the data. The patient class consist of 41% inpatient and 59% outpatient cases. Figure 2 shows how patients were distributed among different age groups.

<table>
<thead>
<tr>
<th>Priority Class</th>
<th>Number of Cases</th>
<th>% of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emergent</td>
<td>1886</td>
<td>2.97</td>
</tr>
<tr>
<td>Immediate</td>
<td>960</td>
<td>1.51</td>
</tr>
<tr>
<td>Organ Donor</td>
<td>50</td>
<td>0.08</td>
</tr>
<tr>
<td>Urgent</td>
<td>1634</td>
<td>2.58</td>
</tr>
<tr>
<td>Elective</td>
<td>58873</td>
<td>92.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ASA Class</th>
<th>Number of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11856</td>
</tr>
<tr>
<td>2</td>
<td>25066</td>
</tr>
<tr>
<td>3</td>
<td>21473</td>
</tr>
</tbody>
</table>

Table 1. Cases by Priority Class

Table 2. Cases by ASA Class
Although surgeon name was available as part of the data, however, this field was not considered as a factor, because a surgeon based model makes the model not to be used for surgeons who just join the practice or those who start performing a new surgery. Data was split with 85 percent for training and the remaining for the test set. The analysis was performed using JMP PRO 10 software. We also note that all time units throughout the paper are in minutes.

3.2. Classification
The dataset shows record of 2000 different procedure codes performed over the period of 39 months. The analysis indicates that procedure code is highly significant factor in both surgery time (ST) and ASP. However, some of the procedure codes are very rare and there are not enough records of these surgery types available. Also there are procedure codes that are very similar in terms of duration. In order to consider the effect of procedure code while fixing some of the issues with rare cases, we decide to define a new variable based on the grouped procedure codes. In order to do that, we use data mining technique, particularly classification to categorize the 2000 different procedure types in new groups. Classification or decision tree is a platform that recursively partitions data according to a relationship between certain independent and dependent variables, creating a tree of partitions. It finds a set of cuts or groupings of independent variables that best predict a dependent variable. These splits of data are done recursively forming a tree of decisions until the desired fit is reached. There are several heuristic algorithms used to build classification and decision trees [21,22]. Some of these algorithms are ID3, and CART [21]; in this study we used CART. The proposed work uses classification to group different procedure codes based on their length of OR time. This allows reduction of the number of categories of procedure codes. This reduction needs to be validated by using cross-validation. We use a 10 fold cross validation to validate the categorization at 95% confidence. The new categorization then can replace the variable procedure code in our regression model. This new variable called adjusted procedure code. The classification method categorizes 2000 procedure types to 49 distinct groups. Therefore, nominal variable adjusted procedure code has 49 levels instead of the initial 2000 levels. The $R^2$ value for the classification is 0.65. We apply Tukey Kramer HSD test with $\alpha=.05$ to test that these categories represent statistically different groups in terms of mean procedure times. The results of this test are very lengthy and therefore are omitted.

3.3. Prediction
Once the procedure category variable is created, regression models are developed to predict the duration. A multi regression is a regression with more than one independent variable or factor and is one of the common methods of prediction. The two of the most common techniques used in multiple regressions are least square regression and stepwise regression. These methods are applied to predict surgery duration. We note that the stepwise regression and linear least square regression are very similar in nature. The difference is mainly in the way significant variables selected. For stepwise regression we use the combination of backward elimination and forward selection.
According to the Gauss–Markov theorem, there are a few conditions to be satisfied such that the least square estimator will be the best linear unbiased estimator. We notice that validation of these conditions is often ignored in many previous studies, resulting in poor or invalid outcomes. In practice, there are rarely situations that all these conditions hold; therefore, there is often need for adjustments and changes in the data that can help to satisfy these. These conditions are:

1. Relationship between dependent and independent variables should be linear
2. The residuals are normally distributed with mean close to zero
3. There is no heteroscedasticity which means that residuals have a constant variance
4. There is no autocorrelation, that is successive residuals are not correlated
5. There is no multi-collinearity

In preparation for use of regression model, all above assumptions have been verified. Due to the length of results related to these assumptions, here only the method that applied to test each of the assumptions has been listed in Table 3.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Method to Verify</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Looking for un even distribution of standardized residuals as function of standardized predicted value around zero horizontal line</td>
</tr>
<tr>
<td>2</td>
<td>Distribution of residuals</td>
</tr>
<tr>
<td>3</td>
<td>Applying white’s test</td>
</tr>
<tr>
<td>4</td>
<td>Applying Durbin Watson test</td>
</tr>
<tr>
<td>5</td>
<td>Calculating correlation between response and the factors (for age) and visual examination of graph of response by factor (for factors other than age)</td>
</tr>
</tbody>
</table>

Also regression factor statistical analysis indicates that all main factors shown in Table 4 are significant factors for least square prediction model with p-value<0.0001. Although we admit that these are not the only factors influencing OR times however the statistical results show that all of these factors significantly affect OR times. For the stepwise regression we not only consider these main factors but also the two level interactions of these factors to start the stepwise regression. The stepwise method however enters only those factors that have a significant impact on the result. We also tested the stepwise with three levels of interaction; however, no improvement has been reported from this model compared to the model with only two levels of interaction. Therefore, in result section only the stepwise method with two levels of interaction has been discussed.

<table>
<thead>
<tr>
<th>Independent Variables (Factors)</th>
<th>Type</th>
<th>Number of Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority Class</td>
<td>Nominal</td>
<td>5</td>
</tr>
<tr>
<td>Procedure Category</td>
<td>Nominal</td>
<td>49</td>
</tr>
<tr>
<td>ASA Class</td>
<td>Nominal</td>
<td>6</td>
</tr>
<tr>
<td>Age</td>
<td>Continuous</td>
<td>-</td>
</tr>
<tr>
<td>Patient Class</td>
<td>Nominal</td>
<td>2</td>
</tr>
<tr>
<td>Specialty</td>
<td>Nominal</td>
<td>30</td>
</tr>
</tbody>
</table>

As we are predicting duration of times, we also need to make sure that the values reported as output of the regression model are greater than zero. Plotting the distribution of duration of surgeries, it could be seen that the distribution is very much skewed to the left. We expect this to cause the prediction values to occasionally get values of zero or even negative. To prevent that, we use log transformation as predictor in regression model. This
transformation has been widely used in literature to prevent the output of time predictions from falling to negative numbers. After applying the regression model, inverse transformation needs to be applied before statistical results are gathered. Figure 3 shows the distribution of case durations in train set before and after transformation.

![Figure 3. Case Distribution Before and After Transformation](image)

4. Results
The results of the three predictions (ASP, LIN, and STEP) are compared against each other. Multiple performance measures such as $R^2$, RASE, and AAE which are coefficient of determination, root square average error, and average absolute error respectively, are reported for both train and test sets (due to respective strengths [23, 24]). Table 5 shows the comparison statistics of the three methods. As can be seen, STEP prediction is better than LIN, and each outperforms ASP.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Train Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>RASE</td>
</tr>
<tr>
<td>ASP</td>
<td>0.6375</td>
<td>59.26</td>
</tr>
<tr>
<td>LIN</td>
<td>0.6598</td>
<td>57.4116</td>
</tr>
<tr>
<td>STEP</td>
<td>0.6838</td>
<td>55.3505</td>
</tr>
</tbody>
</table>

Table 6 shows the comparison of the three prediction values by specialty, for most frequent specialties with at least 100 cases. STEP is the best model for the most frequent specialties (orthopedics and general surgeries, surgical oncology). LIN is better for urology, ophthalmology, thoracic, vascular, GYN oncology, and gynecology. ASP outperforms both STEP and LIN for otolaryngology, plastic, and GYN oncology procedures. Neuro surgery, obstetrics, and acute care procedures have mixed results. For Psychosurgery cases, the $R^2$ values for ASP and LIN are negative (these values for psychology is not shown in table as this specialty did not have many cases), suggesting that the mean of the group is a better representative of the predicted values. Even though the $R^2$ values for STEP is positive (yet very small), none of the prediction models can accurately estimate the procedure durations of this specialty. The t-test performed with alpha=0.05, suggests that LIN and STEP are significantly different from ASP in terms of the mean residuals.
<table>
<thead>
<tr>
<th>Specialty</th>
<th>Predictor</th>
<th>R-Square</th>
<th>RMSE</th>
<th>MAE</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthopedics</td>
<td>ASP</td>
<td>0.5257</td>
<td>57.9685</td>
<td>39.9777</td>
<td>34.2%</td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.5755</td>
<td>54.8451</td>
<td>36.1095</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.5828</td>
<td>54.3689</td>
<td>35.9546</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>ASP</td>
<td>0.3729</td>
<td>55.0437</td>
<td>37.3926</td>
<td>14.6%</td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.4697</td>
<td>50.616</td>
<td>33.7518</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.4965</td>
<td>49.3186</td>
<td>32.7425</td>
<td></td>
</tr>
<tr>
<td>Otolaryngology</td>
<td>ASP</td>
<td>0.7319</td>
<td>51.7225</td>
<td>32.9076</td>
<td>11.3%</td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.7126</td>
<td>53.5506</td>
<td>32.2944</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.7189</td>
<td>52.9621</td>
<td>32.0055</td>
<td></td>
</tr>
<tr>
<td>Urology</td>
<td>ASP</td>
<td>0.7272</td>
<td>46.7483</td>
<td>30.5251</td>
<td>7.8%</td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.7733</td>
<td>42.614</td>
<td>26.7203</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.7709</td>
<td>42.8332</td>
<td>26.9698</td>
<td></td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>ASP</td>
<td>0.2313</td>
<td>35.9271</td>
<td>22.6226</td>
<td>6.3%</td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.5206</td>
<td>28.3732</td>
<td>17.5264</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.4968</td>
<td>29.0675</td>
<td>17.6161</td>
<td></td>
</tr>
<tr>
<td>Neuro Surgery</td>
<td>ASP</td>
<td>0.5946</td>
<td>81.0404</td>
<td>54.9013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.5947</td>
<td>81.0349</td>
<td>52.2301</td>
<td>5.1%</td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.5893</td>
<td>81.5648</td>
<td>52.1195</td>
<td></td>
</tr>
<tr>
<td>Surgical Oncology</td>
<td>ASP</td>
<td>0.6781</td>
<td>54.9955</td>
<td>32.3168</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.691</td>
<td>53.8794</td>
<td>31.6966</td>
<td>4.6%</td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.7119</td>
<td>52.0292</td>
<td>30.4564</td>
<td></td>
</tr>
<tr>
<td>Plastic</td>
<td>ASP</td>
<td>0.7299</td>
<td>78.7518</td>
<td>54.2446</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.6102</td>
<td>94.609</td>
<td>58.9855</td>
<td>4.0%</td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.6412</td>
<td>90.7662</td>
<td>58.5443</td>
<td></td>
</tr>
<tr>
<td>Thoracic</td>
<td>ASP</td>
<td>0.5998</td>
<td>99.1005</td>
<td>66.6381</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.7161</td>
<td>83.4766</td>
<td>59.6474</td>
<td>2.8%</td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.6775</td>
<td>88.958</td>
<td>60.5216</td>
<td></td>
</tr>
<tr>
<td>Vascular</td>
<td>ASP</td>
<td>0.6072</td>
<td>60.2528</td>
<td>40.1038</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.7247</td>
<td>50.4412</td>
<td>35.8458</td>
<td>2.0%</td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.7008</td>
<td>52.5839</td>
<td>37.4437</td>
<td></td>
</tr>
<tr>
<td>Obstetrics</td>
<td>ASP</td>
<td>0.6049</td>
<td>34.3594</td>
<td>26.9036</td>
<td>1.8%</td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.6995</td>
<td>29.9654</td>
<td>20.1546</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.673</td>
<td>31.2549</td>
<td>19.1835</td>
<td></td>
</tr>
<tr>
<td>Psychosurgery</td>
<td>ASP</td>
<td>-2.2645</td>
<td>7.9547</td>
<td>4.5975</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>-0.0481</td>
<td>4.5072</td>
<td>3.3032</td>
<td>1.7%</td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.0009</td>
<td>4.4006</td>
<td>3.2347</td>
<td></td>
</tr>
<tr>
<td>Acute Care Surgery</td>
<td>ASP</td>
<td>0.2728</td>
<td>51.8505</td>
<td>38.2971</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.3784</td>
<td>47.9386</td>
<td>31.887</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.3583</td>
<td>48.7057</td>
<td>31.8369</td>
<td></td>
</tr>
<tr>
<td>GYN Oncology</td>
<td>ASP</td>
<td>0.6277</td>
<td>56.7332</td>
<td>35.9823</td>
<td>1.2%</td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.6227</td>
<td>57.1121</td>
<td>36.812</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STEP</td>
<td>0.6144</td>
<td>57.7381</td>
<td>36.4322</td>
<td></td>
</tr>
<tr>
<td>Gynecology</td>
<td>ASP</td>
<td>0.5071</td>
<td>52.1249</td>
<td>37.6535</td>
<td>1.1%</td>
</tr>
<tr>
<td></td>
<td>LIN</td>
<td>0.6241</td>
<td>45.5197</td>
<td>30.9478</td>
<td></td>
</tr>
</tbody>
</table>
5. Discussion
Both LIN and STEP show moderate improvement over ASP. This is not surprising since ASP prediction is often adjusted by the scheduler and/or the surgeon. These adjustments are typically based on intuitive consideration of patient characteristics and clinical factors and surgery complexity (12). There is no indication in the data set of any of these variables however about how often the moving average has been adjusted and by how much to determine reported scheduled duration. We believe that additional factors such as those considered by scheduler to adjust the value of the moving average if available can potentially make LIN and STEP even more accurate. The STEP regression model includes two-level interactions (inclusion of three-level interactions gives only slight improvement). The performance of LIN and STEP is dependent on the choice of independent variables and the size of the data set. While having additional factors and interaction of factors can potentially increase $R^2$ value in the training set indicating a better model, corresponding performance in the test set needs to be evaluated.

The independent variables considered in this study are general factors available across all specialties, and hence ensures ease of scalability. If more clinical factors are added to the model, then the interaction of factors will add more value to the accuracy of the prediction. These clinical factors however may not be available in structured data format. In some situations these factors could be observed by applying text mining on pre and post diagnostic notes if these notes are recorded electronically in unstructured form. However, in some situations such data is not available electronically. In such case there is need for manual observation of factor from notes by reading hand written charts by experts. We also note that each specialty has several meaningful variables which are specific to their practice, to add all these factors into a single model, care needs to be taken since this will create an uneven spread of factors among data due to the fact that some practices perform more surgeries than others. For instance, our data shows an indication that there is more of general and orthopedic surgery compared with other specialties; this however should not be used against accuracy of the prediction of duration of specialties with lesser number of cases. Further, individual specialties tend to show variation in performance based on the type of regression. STEP performs best with fewer specialties but those that account for more than half of the procedures. LIN performs best with specialties that account for a third of the procedures. Psychosurgery as a specialty is not predicted well by any of the models, and hence needs further investigation.

6. Conclusion
Prediction of OR times is very important as these times are used to assign time and day of the surgery. Accurate predictions are necessary to prevent over- and under-utilization. We proposed and evaluated a hybrid method with two steps 1) creation of a new variable to categorize procedures across all specialties and 2) development of regression models to predict procedure duration using the procedure category variable from the first step along with other factors. Evaluation shows that both regression models (LIN and STEP) result in better predictions compared to the current state-of-the-practice. The hybrid method with STEP regression gives a better prediction for orthopedics and general surgeries and surgical oncology specialties, which constitute more than half of the procedures. The proposed hybrid method can effectively deal with the heterogeneity problem, and further improvements can be obtained through inclusion of additional clinical factors.

References

| STEP   | 0.6195 | 45.7996 | 31.8781 |
17. Pandit JJ, Tavare A. Using mean duration and variation of procedure times to plan a list of surgical operations to fit into the scheduled list time. European Journal of Anaesthesiology (EJA). 2011;28(7):493-501.
24. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Manuscript prepared for Geosci. Model Dev. Discuss. with version 4.1 of the LATEX class copernicus discussions.cls. ed2014.
Medical Inpatient Journey Modeling and Clustering: A Bayesian Hidden Markov Model Based Approach
Zhengxing Huang, PhD¹, Wei Dong, PhD², Fei Wang, PhD³, Huilong Duan, PhD¹
¹College of Biomedical Engineering and Instrument Science, Zhejiang University
²Department of Cardiology, Chinese PLA General Hospital
³Department of Computer Science and Engineering, University of Connecticut

Abstract
Modeling and clustering medical inpatient journeys is useful to healthcare organizations for a number of reasons including inpatient journey reorganization in a more convenient way for understanding and browsing, etc. In this study, we present a probabilistic model-based approach to model and cluster medical inpatient journeys. Specifically, we exploit a Bayesian Hidden Markov Model based approach to transform medical inpatient journeys into a probabilistic space, which can be seen as a richer representation of inpatient journeys to be clustered. Then, using hierarchical clustering on the matrix of similarities, inpatient journeys can be clustered into different categories w.r.t their clinical and temporal characteristics. We evaluated the proposed approach on a real clinical data set pertaining to the unstable angina treatment process. The experimental results reveal that our method can identify and model latent treatment topics underlying in personalized inpatient journeys, and yield impressive clustering quality.

1. Introduction
In recent years, healthcare process management is increasingly being used in healthcare organizations to provide standardized and normalized health services in medical inpatient journeys [1, 2]. Different from common business processes in commercial and industrial environments, healthcare processes are highly dynamic, context sensitive, event driven, and knowledge intensive such that they often bear no relation to the ideal as envisaged by the designers of healthcare processes [3]. To this end, healthcare organizations need to analyze and improve healthcare processes continuously [4].

Regarding healthcare process analysis and improvement, process mining, as a valuable set of techniques in business process management, has achieved emerging attention in clinical settings [3, 5, 6]. Process mining techniques use event logs to record business process execution information, to mine the actual behaviors in business processes, and to discover business process models from event logs [7]. Shifting to clinical settings, process mining can be an objective way of analyzing healthcare processes as it is not biased by perceptions or normative behaviors [5]. Applying process mining in clinical settings, non-trivial information about healthcare processes can be extracted from electronic medical records (EMRs) that contains the execution results of inpatient journeys, such as latent treatment patterns [3], treatment performance metrics [8], and performance characteristics [9], etc.

However, the diversity of treatment behaviors in healthcare processes is far higher than that of common business processes [5]. Healthcare processes are typically dynamic, complex, and loosely-structured, such that traditional process mining techniques have many problems and challenges when applied for healthcare process analysis and improvement [3, 10]. In fact, the diversity of healthcare processes, i.e. each inpatient journey has different kind of treatment events as well as difficult sequences of treatment events, causes the mining results often complicated and difficult to understand.

As a fundamental research problem, clustering actual medical inpatient journeys plays an important role in assisting healthcare process analysis and improvement. Indeed, a key step for healthcare process analysis is to cluster similar inpatient journeys into homogeneous subsets (clusters) [5, 11]. This helps clinical analysts locate treatment information of interest and capture an overview of healthcare processes easily and quickly. It must mention that, in contrast to static health data clustering, the clustering of medical inpatient journeys needs to be performed for each type of healthcare processes and be limited to the number of patients following a particular healthcare process protocol (e.g., a specific clinical guideline or pathway to a specific disease, etc.). This adds extra requirements to clustering, i.e., the clustering algorithm should group similar inpatient journeys together, and separate relevant
inpatient journeys from irrelevant ones in order to support medical staff to efficiently understand and browse these journeys for the further tasks on healthcare process analysis and improvement.

The requirements above in general introduce several challenges to modeling and clustering of medical inpatient journeys. In contrast to structured business process execution traces, an inpatient journey is described as a series of clinical epochs/stages in the patient’s hospitalization, and each epoch consists of a set of treatment events which may occur arbitrarily without a particular order [12]. Thus, one has to incorporate such loosely-structured treatment event sequence in the clustering process for good similarity measure. To the best of our knowledge, most of the approaches developed so far try to modify the existing algorithms to handle sequential business process data [6, 13, 14]. In order to have benefits, the data sources should be in fine structures. Unfortunately, such types of data sources are not available or rich enough for healthcare processes [5, 6].

In consideration of informal structures and the modelling of non-stationary data on inpatient journeys, we present a new approach for modeling and clustering medical inpatient journeys in this article. The proposed approach combines the model-based method provided by a well-known Bayesian Hidden Markov Model (B-HMM) [15] and a hierarchical clustering procedure. More specifically, we develop a clustering process that incorporates the medical behavior dependency observed in inpatient journeys and transforms inpatient journeys into a probabilistic space, where a symmetric similarity between inpatient journeys can be measured. Then, we resort to hierarchical clustering on the matrix of similarity in order to cluster the data sequences into homogenous groups according to both their characteristics and dynamics. In this sense, our approach provides an intuitive organization of the inpatient journey repository. Experiments on a real clinical data set collected from a Chinese hospital show that our proposal outperforms traditional approaches on medical inpatient journey modeling and clustering.

2. Methods

This section introduces a probabilistic model-based approach for medical inpatient journey modeling and clustering. The proposed approach combines both model-based and hierarchical clustering procedures, as shown in Figure 1. We first introduce the basic concepts and notations.

![Figure 1](image)

**Figure 1.** The proposed approach for medical inpatient journey clustering.

2.1 Representation of a Medical Inpatient Journey

In this study, we make an assumption that treatment events in medical inpatient journeys are regularly recorded in EMRs, which effectively reflects real executing conditions in inpatient journeys. Each treatment event refers to a well-defined step in inpatient journeys. Additional information such as the occurring time-stamp of the event is used in this study. To explain the kind of input needed for the proposed approach, we first define the following concepts.

Let $A$ be a finite set of treatment event identifiers (clinical terms describing activities), and $T$ the time domain set (set of time point primitives). A treatment event $e$ is a pair $e = (a, t)$ where $a \in A$ and $t \in T$. We denote by $E = A \times T$ the set of all valid events of a particular domain. Note that treatment events could be characterized by various properties, e.g., an event has an occurring time stamp, it corresponds to an event type, it is executed for a particular patient, has associated cost, etc. We do not impose a specific set of properties, however, given the focus of this study, we assume that the event type and occurring time stamp of the event are present. For convenience, let $e, a$ and $e, t$ denote the event type and occurring time stamp of $e$, respectively. For example, $e = (admission, 1)$ is a treatment event where $e, a = admission$ is the admission type, and $e, t = 1$ is the occurring time of the event. A medical inpatient journey is a series of treatment events $\sigma = (e_1, e_2, \ldots, e_n)$, observed on a particular patient in his/her hospitalization.

For a particular medical inpatient journey $\sigma = (e_1, e_2, \ldots, e_n)$, we have $e_1, t \leq e_2, t \leq \cdots \leq e_n, t$. In general, an inpatient journey consists of different categories of treatment events, and certain temporal dependencies exist between the events for a particular inpatient journey. Figure 2 depicts examples of medical inpatient journeys.

2.2 Medical inpatient journey modeling
In this section, we present a probabilistic model to recognize medical inpatient journeys by considering the sequential and dependency characteristics of treatment information in the journeys. As we mentioned above, inpatient journeys are dynamic, flexible, and loosely-structured. Although there are temporal dependencies between some critical events in different clinical epochs of inpatient journeys, treatment events in one time epoch might occur arbitrarily without a strict order. In this sense, we segment an inpatient journey $\sigma$ as a series of $M$ clinical epochs $\sigma = \{\sigma(1), \sigma(2), \ldots, \sigma(M)\}$, and each epoch has a specific time duration, e.g., a hospitalization day, etc. For example, inpatient journey $\sigma_2$ shown in Figure 2 consists of 5 epochs sequentially, in which each epoch records typical treatment events occurring on a particular hospitalization day. For example, the first clinical epoch of $\sigma_1$ consists of 26 treatment events, i.e., $\sigma_1(1) = \{(A01,1), (A11,1), (A21,1), \ldots\}$, and the last epoch of $\sigma_1$ has one treatment event, i.e., $\sigma_1(5) = \{(A17,5)\}$.

In addition, we assume that each clinical epoch of a particular inpatient trajectory has a specific treatment topic (e.g., “Admission”, “Prepare Surgery”, “Surgery”, “Post-surgery recovery”, etc.). Formally, we use $z_{\sigma(t)}(t)$ to denote the underlying treatment topic in the $t$th epoch of a particular inpatient journey $\sigma$. Thus the objective of inpatient journey modeling is to identify the latent treatment topics and their transitions in a particular inpatient journey.

To this end, we propose a Bayesian Hidden Markov Model (B-HMM) [15] based medical inpatient journey model (IJM) to identify inpatient journeys. The proposed IJM has the structure of a standard HMM that contains symmetric Dirichlet priors over the transition and emission distributions for modeling the sequential treatment information in an individual inpatient journey. Formally, given an inpatient journey $\sigma = \{\sigma(1), \sigma(2), \ldots, \sigma(M)\}$ with a series of $M$ clinical epochs, the dependency relationships of IJM are represented as follows:

$$z_{\sigma(t)}(t) \sim \text{Multinomial}(\theta_{z_{\sigma(t-1)}}(t-1)) \quad (1); \quad e_{t} | z_{\sigma(t)}(t), \Phi \sim \text{Multinomial}(\phi_{z_{\sigma(t)}}(t)) \quad (2)$$

$$\theta_{z_{\sigma(t-1)}}(t-1) | \alpha \sim \text{Dirichlet}(\alpha) \quad (3); \quad \phi_{z_{\sigma(t)}}(t) | \beta \sim \text{Dirichlet}(\beta) \quad (4)$$

Where $z_{\sigma(t)}(t) \sim \text{Multinomial}(\theta_{z_{\sigma(t-1)}}(t-1))$ means $z_{\sigma(t)}(t)$ follows multinomial distribution $\text{Multinomial}(\theta_{z_{\sigma(t-1)}}(t-1))$, $\theta_{z_{\sigma(t-1)}}(t-1)$ is the topic transition distribution over the $t$th clinical epoch $\sigma(t)$ when the treatment topic of the previous epoch $\sigma(t-1)$ is $z_{\sigma(t-1)}$. $z_{\sigma(t)}(t)$ indicates the treatment topic in $\sigma(t)$. $e_{t}$ is the $t$th treatment event recorded in $\sigma(t)$, and $\phi_{z_{\sigma(t)}}(t)$ is the emission distribution of treatment events in $\sigma(t)$. Particularly, $\alpha$ and $\beta$ follow the Dirichlet distribution with parameters $\alpha$ and $\beta$. Figure 3 shows the graphical representation of IJM.

In summary, the processed IJM assumes the following generative process for an inpatient journey $\sigma$:

1. Draw treatment topic proportions $\theta_{z} \sim \text{Dirichlet}(\alpha)$
2. For each treatment topic $z = 1, \ldots, K$, draw treatment event probability $\phi_{z} \sim \text{Dirichlet}(\beta)$
3. For each clinical epoch $t = 1, \ldots, M$ of $\sigma$:
   3.1 Draw treatment topic $z_{\sigma(t)}$ from $\theta_{z_{\sigma(t)}}$ w.r.t the previous treatment topic $z_{\sigma(t-1)}$, $z_{\sigma(t)} \sim P(z_{\sigma(t)} | z_{\sigma(t-1)}, \theta_{z_{\sigma(t)}})$
   3.2 For each treatment event $e_{t} \in \sigma(t)$, $t = 1, \ldots, |\sigma(t)|$, draw $e_{t} \sim \text{Multinomial}(\phi_{z_{\sigma(t)}}(t))$.

Given the generative process of IJM, we can calculate the joint distribution of all observations and hidden variables in IJM by the following equation:

$$P(z_{\sigma(t)}, \sigma(t), \theta, \Phi | z_{\sigma(t-1)}, \alpha, \beta) = P(\theta | \alpha)P(z_{\sigma(t)} | z_{\sigma(t-1)}, \theta)P(\Phi | \beta) \left( \prod_{i=1}^{\sigma(t)} P(e_{t} | z_{\sigma(t)}(t), \Phi) \right) \quad (5)$$

Therefore, the likelihood of a medical inpatient journey $\sigma$ can be calculated as follows:

$$L(\sigma) = \int \prod_{t=1}^{K} P(\theta_{z} | \alpha) \prod_{t=1}^{M} P(e_{t} | z_{\sigma(t)}, \Phi_{z_{\sigma(t)}}(t)) \, d\theta \int \prod_{t=1}^{K} P(\phi_{z} | \beta) \prod_{t=1}^{M} \prod_{i=1}^{\sigma(t)} P(e_{t} | z_{\sigma(t)}, \Phi_{z_{\sigma(t)}}(t)) \, d\Phi \quad (6)$$

We developed a Gibbs sampling approach to get the to maximize the likelihood in Equation (6), with the time complexity for a particular inpatient journey $\sigma$ being $O(LKM)$, where $L$ is the number of iterations, $K$ is the number of treatment topics, and $M$ is the number of clinical epochs in $\sigma$.

2.3 Medical inpatient journey clustering

---

4 In this study, we set the time range of each clinical epoch in an inpatient journey $\sigma$ as a hospitalization day, which records a set of treatment events observed on a particular day in the patient’s length of stay.
The second step of our hybrid clustering method exploits the information provided by the probabilistic space obtained in the first step to define the clusters of medical inpatient journeys characterized by similar treatment behaviors. Based on the generated IJM models of inpatient journeys, we measure the similarity between the journeys. Specifically, we use Monte Carlo sampling to compare two IJM models. Formally, let $M_i$ and $M_j$ be the learned IJM for inpatient journeys $\sigma_i$ and $\sigma_j$, respectively. The similarity between any two IJM models is defined as

$$ Sim(M_i, M_j) = \frac{\log p(\sigma_i|M_j) + \log p(\sigma_j|M_i)}{2} $$

(13)

<table>
<thead>
<tr>
<th>A0: Admission</th>
<th>A14: Routine blood test</th>
<th>A28: Anti-arrhythmic examination</th>
<th>A42: Diuretic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1: ECG examination</td>
<td>A15: Transfer</td>
<td>A29: Local anesthesia</td>
<td>A43: Punction</td>
</tr>
<tr>
<td>A2: Ultrasound examination</td>
<td>A16: Coronary angiography</td>
<td>A30: Diabetes examination</td>
<td>A44: Peripheral vasodilator</td>
</tr>
<tr>
<td>A3: Cardiovascular treatment</td>
<td>A17: Discharge</td>
<td>A31: Thyroid function tests</td>
<td>A45: Plasma concentration</td>
</tr>
<tr>
<td>A4: ß-adrenergic receptor blockers</td>
<td>A18: Radiation</td>
<td>A32: Anesthesia</td>
<td>A46: CT examination</td>
</tr>
<tr>
<td>A5: Antithrombotic treatment</td>
<td>A19: Oesophagus blood test</td>
<td>A33: HE receptor antagonist</td>
<td>A47: Tumor markers check</td>
</tr>
<tr>
<td>A6: Anticoagulant treatment</td>
<td>A20: Stent implantation</td>
<td>A34: Renal arteriography</td>
<td>A48: First-level care</td>
</tr>
<tr>
<td>A7: Antiplatelet treatment</td>
<td>A21: PTCA</td>
<td>A35: Angiotensin receptor antagonists</td>
<td>A49: Routine care</td>
</tr>
<tr>
<td>A8: Biochemical examination</td>
<td>A22: Angiotensin-converting enzyme inhibitors</td>
<td>A36: Antihypertensive treatment</td>
<td>A50: Specific meal</td>
</tr>
<tr>
<td>A10: Routine stool test</td>
<td>A24: Coagulation examination</td>
<td>A38: Typhon T</td>
<td>A52: Oxygen inhalation</td>
</tr>
<tr>
<td>A13: Blood serum test</td>
<td>A27: Calcium channel blockers</td>
<td>A41: Calcium regulation</td>
<td></td>
</tr>
</tbody>
</table>

Where $\log p(\sigma_i|M_j)$ and $\log p(\sigma_j|M_i)$ are a measure of how well model $M_j$ matches observations generated by model $M_i$, relative to how well $M_i$ matches the observations generated by itself. $\log p(\sigma_i|M_j)$ and $\log p(\sigma_j|M_i)$ are in the same spirit and make the similarity $Sim(M_i, M_j)$ symmetric. Equation (13) can be rewritten in terms of inpatient journeys as

Figure 2. Examples of medical inpatient journeys
\[
Sim(\sigma_i, \sigma_j) = Sim(M_i, M_j) = \frac{1}{2} \sum_{t=1}^{M_j} \log P(\sigma_i(t) | M_i) + \frac{1}{2} \sum_{t=1}^{M_i} \log P(\sigma_j(t) | M_j)
\] (14)

Where \(\sigma_i(t)\) is the \(t\)-th clinical epoch in \(\sigma_i\), \(\sigma_j(t)\) is the \(t\)-th clinical epoch in \(\sigma_j\), and the log-likelihood for each inpatient journey given the IJM can be obtained from Equation (8). This similarity measure is well suited to large collection of inpatient journeys, as it only requires the storage of the IJM parameters for each piece rather than the original patient’s EMR data itself.

In order to make no assumption on the number of medical inpatient journey clusters that are to be extracted, we use hierarchical clustering, which builds a hierarchy of clusters rather than treating all clusters as distinct, equal entities, such as in K-Means clustering.

A reasonable similarity measure \(Sim(\sigma_i, \sigma_j)\) for medical inpatient journeys \(\sigma_i\) and \(\sigma_j\) is critical for inpatient journey clustering. In Equation (15), we have presented how to measure \(Sim(\sigma_i, \sigma_j)\) between inpatient journeys \(\sigma_i\) and \(\sigma_j\), which defines the similarity matrix used for grouping the collected journeys into a tree of clusters through the hierarchical clustering procedure [16]; the dendrogram is achieved starting from \(C\) clusters, one for each inpatient journey \(\sigma\), and iteratively aggregating pairs of clusters until one single clusters is obtained. This agglomerative strategy is defined using complete linkage, i.e., similarity between clusters is measured on the basis of the similarity between the two furthest data points in two clusters. The hierarchical clustering allows us to easily obtain an informative data structure without having to specify a priori the number of clusters.

3. Experiments

In this case study, a collection of EMRs consisting of 9944 medical inpatient journeys following the unstable angina treatment process (from 2004 to 2013) was extracted from hospital information systems of Chinese PLA General Hospital to demonstrate the feasibility of the proposed approach. The collected data-set have 704004 treatment events within 606 event types. The average length of stay (LOS) recorded in the collection of EMRs is 8.30 days, which some inpatient journeys take a very short time, e.g., only 1 day in hospital, and other trajectories take much longer, e.g., more than 3 months in the hospital, which implicitly indicates the diversity of inpatient journeys in the unstable angina treatment process.

The case study was performed in the Cardiology Department at the Chinese PLA General Hospital. Prior approval was obtained from the data protection committee of the hospital to conduct the study. We state that the patient data was anonymized in this study and in this paper. All experiments were performed on a Lenovo Compatible PC with an Intel Pentium IV CPU 2.8 GHz, 4G byte main memory running on Microsoft Windows 8.1. The algorithm was implemented using Microsoft C#.

To evaluate the proposed IJM, we also developed a traditional sequence alignment based inpatient journey clustering method, and a variation HMM to model and cluster inpatient journeys. For the variation HMM, it can label the \(t\)-th clinical epoch of a particular inpatient journey \(\sigma\) by

\[
P(z_t = z | \sigma(t)) \propto P(\sigma(t) | z_t \equiv z)P(z_t | z_{t-1}) \propto P(z_t | z_{t-1}) \Pi_{i=1}^{\alpha(t)}P(e_i | z_t \equiv z)
\] (15)

Based on the suggestion of our clinical collaborators, the number of latent treatment topics for both IJM and HMM was chosen as 5, which is the general number of clinical stages of the unstable angina treatment process. For the other parameters of the proposed IJM, we used the following values: \(\alpha = 0.1, \beta = 0.01\), and the number of Gibbs iterations \(L=1000\).

**Modeling performance.** In this subsection, we evaluate the proposed IJM on modeling medical inpatient journeys. To this end, we randomly picked up 10 sample inpatient journeys from the collected dataset, as shown in Figure 2. Since the whole time period of the collected EMRs is across 10 years (i.e., from 2004 to 2013), we randomly selected one piece of EMRs in the dataset segment of a particular year. Each selected sample records typical
treatment behaviors occurred in a particular inpatient journey.

To investigate the problem of how to know the learned treatment topics are meaningful or not for a given medical inpatient journey, we adopted a hypothesis testing in which we used the derived treatment topics to describe each clinical epoch of an inpatient journey, and then conducted statistical tests to know whether the derived treatment topics from IJM can better represent the treatment topical information of a particular clinical epoch than that of the variation HMM on a given inpatient journey. The null hypothesis assumes that “both IJM and HMM have no difference in treatment topical representation in inpatient journeys”. The process to test the significance of violating the null hypothesis is given as follows:

(1) Firstly, we labeled each clinical epoch $\sigma(t)$, $(1 \leq t \leq M)$ of a sample inpatient journey $\sigma$ with the recognized treatment topic $z$ which is derived from either IJM or HMM. In particular, we chose the treatment event types $e$, $a$ with $P(e, a | z) > 0.01$ to represent the derived treatment topic $z$.

(2) Secondly, we asked 3 clinicians from the Cardiology department of the hospital to evaluate that to what extend the derived treatment topic correctly represents the actual treatment information in each clinical epoch of a sample inpatient journey. To ensure the evaluation quality, we did not inform evaluators that a given learnt treatment topic is learned by which model. The answer was given on a particular label, i.e., “Bad”, “Fair” or “Good” (i.e., “Bad” represents “does not represent at all”, “Fair” represents “fairly represent the treatment information in a specific clinical epoch”, and “Good” represents “perfectly represent the treatment information in a specific clinical epoch”, respectively). Note that we obtained the evaluation result of 3 evaluators based on a major voting strategy.

Figure 4 shows the results of human judgment for both IJM and HMM. From Figure 4, we can observe that the proposed IJM outperforms HMM in terms of “Good” cases and positive cases (“Good” + “Fair”), which indicates that treatment topics learned by our IJM is more reasonable because our model can represent loosely-structured inpatient journeys into the learning process.

(3) We performed both cohen’s Kappa test and $t$-test on the human evaluation results for both IJM and HMM. (i) The Kappa test is used to calculate inter-judge reliability between IJM and HMM. For the good cases, the obtained kappa value for Kappa test is 0.053, which is slightly larger than 0 and indicates the poor agreement between IJM and HMM. For the positive cases, the obtained kappa value for Kappa test is 0.167, indicating slight agreement between IJM and HMM. (ii) The $t$-test is used to evaluate the human understandings on the derived treatment topics and their assignments on clinical epochs of medical inpatient journeys. For the good cases, the obtained $t$ statistic is 6.020, which is larger than 3.250 (the confidence value of 99.5%). It indicates that the proposed IJM achieves the better understanding of human evaluators on the discovered treatment topics and their assignments on clinical epochs of inpatient journeys than that of HMM. For the positive cases, the obtained $t$ statistic is 2.756, which is less than 3.250 (the confidence value of 99.5%), and indicates that there are statistically no differences between IJM and HMM on positive cases.

![Figure 4](image.png)

Figure 4. Spherical comparison in terms of human evaluation. (A) Percentage of good cases; (B) Percentage of positive cases.

To illustrate the quality of inpatient journey modeling more intuitively, we take the inpatient journeys $\sigma_1$ and $\sigma_2$ (as shown in Figure 2) as examples. In particular, we manually checked the learned treatment topics with their representative treatment events for $\sigma_1$ and $\sigma_2$. In consideration of treatment topic and their transitions for both examples, clinical evaluators think that all these transitions are reasonable and easy to understand.
For example, the patient with $\sigma_1$ is an aged male patient. He has high-risk level of unstable angina and his LOS is 5 days. With IJM, a typical treatment topic transition pattern for the unstable angina treatment process is generated:

**Topic 1 (Admission, day 1)** → **Topic 2 (Prepare surgery, day 2)** → **Topic 3 (Surgery, day 3)** → **Topic 4 (Post-surgery recovery, day 4)** → **Topic 5 (Discharge, day 5)**

To look insight into the learned treatment topics, clinical evaluators think that most of treatment event types of each learned topic underlying in $\sigma_2$ are reasonable, except for one irrelevant treatment event type in the learned topic “Prepare surgery“ of $\sigma_1$, which is shown in bold in Table 1. As a result, 4 out of 5 clinical epochs are labeled as “Good” and 1 clinical epoch ($t = 2$) is labeled as “Fair”.

<table>
<thead>
<tr>
<th>$\sigma_1$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic 1</strong></td>
<td>Admission, Routine urine test, Blood typing, Routine stool test, Calcium channel blockers, Routine blood test, Blood serum test, Ultrasonography examination, ECG examination, Cardiovacular treatment, Antianginal treatment, Biochemical examination, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Lipid regulation, Anticoagulation treatment, Angiotensin-converting enzyme inhibitors, Angiotensin receptor antagonists, Proton pump inhibitors, Peripheral vasodilator, Thyroid function tests, X-ray, Occult blood test, Coagulation examination, Radiographic examination, Blood sugar regulation</td>
</tr>
<tr>
<td><strong>Topic 2</strong></td>
<td>Cardiovascular treatment, Anticoagulation treatment, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Antianginal treatment, <strong>Oxygen inhalation</strong>, Proton pump inhibitors, Lipid regulation, Angiotensin-converting enzyme inhibitors, Calcium channel blockers, X-ray, Angiotensin receptor antagonists, Peripheral vasodilator</td>
</tr>
<tr>
<td><strong>Topic 3</strong></td>
<td>Cardiovascular treatment, Anticoagulation treatment, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Antianginal treatment, Lipid regulation, Angiotensin-converting enzyme inhibitors, Anesthesia, Coronary angiography, Stent implantation, Proton pump inhibitors, Calcium channel blockers, Angiotensin receptor antagonists, Peripheral vasodilator</td>
</tr>
<tr>
<td><strong>Topic 4</strong></td>
<td>Cardiovascular treatment, Anticoagulation treatment, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Antianginal treatment, Oxygen inhalation, Lipid regulation, Angiotensin-converting enzyme inhibitors, Proton pump inhibitors, Calcium channel blockers, Angiotensin receptor antagonists, Multifunctional monitors</td>
</tr>
<tr>
<td><strong>Topic 5</strong></td>
<td>Antianginal treatment, Discharge</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\sigma_2$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic 1</strong></td>
<td>Admission, Routine stool test, Routine urine test, Blood typing, Routine care, Routine blood test, Blood serum test, Specific meal, Ultrasonography examination, Cardiovascular treatment, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Antianginal treatment, Second-level care, Lipid regulation, Occult blood test, Coagulation examination, <strong>Blood sugar regulation</strong>, Consultation</td>
</tr>
<tr>
<td><strong>Topic 2</strong></td>
<td>Cardiovascular treatment, Anticoagulation treatment, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Antianginal treatment, Lipid regulation, <strong>Blood sugar regulation</strong>, Troponin T</td>
</tr>
<tr>
<td><strong>Topic 3</strong></td>
<td>Cardiovascular treatment, Anticoagulation treatment, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Antianginal treatment, Lipid regulation, <strong>Blood sugar regulation</strong>, <strong>Anti-diabetes treatment</strong></td>
</tr>
<tr>
<td><strong>Topic 4</strong></td>
<td>Cardiovascular treatment, Anticoagulation treatment, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Antianginal treatment, Lipid regulation, Anesthesia, Coronary angiography, Stent implantation, PTCA, Local anesthesia, <strong>Blood sugar regulation</strong>, <strong>Anti-diabetes treatment</strong></td>
</tr>
<tr>
<td><strong>Topic 5</strong></td>
<td>Cardiovascular treatment, Anticoagulation treatment, Antiplatelet treatment, $\beta$-adrenergic receptor blockers, Antianginal treatment, Lipid regulation, Coagulation examination, Discharge, <strong>Anti-diabetes treatment</strong></td>
</tr>
</tbody>
</table>

The patient with $\sigma_2$ is an aged female unstable angina patient. Her LOS is 8 days. This patient also has a common comorbidity of unstable angina, i.e., “Diabetes”, such that anti-diabetes treatments, e.g., “Diabetes check”, “Glucose regulating treatment”, etc., were performed during her LOS (shown in bold in Table 1). In clinical practice, conservative treatments are performed in $\sigma_2$. The proposed IJM generates the following treatment topic transition pattern:

**Topic 1 (Admission, day 1)** → **Topic 2 (day 2)** → **Topic 3 (day 3)** → **Topic 4 (Surgery, day 4)** → **Topic 3 (days 5-7)** → **Topic 5 (Discharge, day 8)**
To look insight into the learned treatment topics of $\sigma_2$, clinical evaluators find that most of treatment event types in the derived topics of $\sigma_2$ are reasonable. In particular, they point out that our approach cannot only discover typical medical interventions on unstable angina treatment and therapy, but also disclose typical medical interventions on Diabetes treatment for the patient, such as “Blood sugar regulation”, and “Anti-diabetes treatment”, etc., as shown in the derived topics of $\sigma_2$. Regarding discovered topics for each clinical epoch, clinical evaluators think it would be better to label the third clinical epoch of $\sigma_2$ as “Topic 2” than “Topic 3”. As a result, clinical evaluators labeled 7 clinical epochs as “Good”, and the other one as “Fair” for $\sigma_2$.

Figure 5. Part of the cluster hierarchies obtained by applying the proposed IJM (A), a variation HMM (B), and sequence alignment (C). Every node represents a cluster and reports the number of traces in the cluster itself, and their average normalized similarity (in brackets).

Clustering performance. The hierarchical clustering on the experimental dataset was performed based on the similarity matrix using Equation (14). Since the similarity measure is the key of clustering techniques, we compared the presented IJM-based similarity measure with the variation HMM-based similarity measure, and the classical sequence-alignment-based (SA) similarity measure [16].
To evaluate the proposed approach in a quantitative manner, we measured the cluster homogeneity, which is a widely used measure for the quality evaluation of clustering [22, 43]. A classical definition of cluster homogeneity is as follows:

\[ H(C) = \frac{\sum_{i,j \in C} \text{sim}(\sigma_i, \sigma_j)}{|C|^2} \]  

(16)

Where \(|C|\) is the number of inpatient journeys in cluster \(C\), and \(\text{sim}(\sigma_i, \sigma_j)\) is the similarity between any two inpatient journeys \(\sigma_i\) and \(\sigma_j\) in \(C\). The higher the homogeneity value, the better the quality of clustering results.

Figure 5 shows parts of the cluster hierarchies we obtained by applying the proposed IJM, HMM, and SA, respectively, on the experimental dataset. The structure of the hierarchies and the content of the resulting clusters are very different between the proposed approach and the baseline methods. As shown in Figure 5(B) and (C) the hierarchies built by both HMM and SA are a bit friable: the root node has a lot of children. It indicates that both HMM and SA resort to a large amount of clusters (117 and 188 for HMM and SA, respectively). In addition, the generated hierarchies by both HMM and SA are very unbalanced: one of these children corresponds to a very big cluster, while the others contain only one or just a few inpatient journeys. In comparison with the benchmark methods, the hierarchy generated by our approach is not sparse, and each node is normally split into few clusters of more comparable dimensions (cc. Figure 5(A)).

In addition, the average homogeneity \(H\) were calculated on the obtained clusters from data to assess its quality [11]. Average cluster homogeneity allows to compare the output of different clustering techniques on the same dataset. Figure 5(A) shows that the proposed approach had an average homogeneity of 0.54. On the other side, using HMM and sequence alignment, they reached an average homogeneity of 0.04 and 0.13, respectively. These indicate that the proposed method outperforms the benchmark methods based on medical inpatient journey clustering.

As shown in Figure 5(A), three clusters in the second level generated by our approach define the cut of the cluster tree that corresponds to the maximum split between clusters, which allow us to obtain some relevant insights on the dynamics of the unstable angina treatment process:

**Table 2.** Clustering results on the unstable angina dataset using IJM. 25 top-ranked treatment activities are listed to refer to clusters at level 2 in the hierarchies. Unique treatment activities of each cluster are marked with bold type.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Antiplatelet treatment, Antianginal treatment, Routine blood test, Anticoagulation treatment, Ultrasonography examination, Lipid regulation, Biochemical examination, Discharge, Routine urine test, Admission, Routine stool test, Blood serum test, ECG examination, <strong>Coronary angiography</strong>, Blood typing, Calcium channel blockers, (\beta)-adrenergic receptor blockers, <strong>Anesthesia</strong>, Coagulation examination, Occult blood test, Angiotensin receptor antagonists, Radiographic examination, <strong>Stent implantation</strong>, Troponin T, <strong>PTCA</strong></td>
</tr>
<tr>
<td>2</td>
<td>Antianginal treatment, Routine blood test, Biochemical examination, Ultrasonography examination, Hypnotic sedative and anxiolytic, <strong>Diuretics</strong>, Antiplatelet treatment, Glucose regulating treatment, ECG examination, Routine urine test, Calcium channel blockers, Routine stool test, Lipid regulation, Anticoagulation treatment, Discharge, Coagulation examination, Admission, Electrolyte regulating treatment, Analysis of blood plasma, Blood serum test, Blood sugar regulation, (\beta)-adrenergic receptor blockers, Occult blood test, Angiotensin receptor antagonists, Radiographic examination</td>
</tr>
<tr>
<td>3</td>
<td>Ultrasonography examination, CT examination, Antianginal treatment, Routine blood test, Biochemical examination, Routine urine test, Routine stool test, Admission, Lipid regulation, Discharge, Coagulation examination, Glucose regulating treatment, Calcium channel blockers, Blood serum test, Antiplatelet treatment, Occult blood test, Radiographic examination, Blood typing, ECG examination, <strong>Peripheral vasodilator</strong>, Anticoagulation treatment, <strong>Tumor markers checks</strong>, Analysis of blood plasma, <strong>Consultation, Transfer</strong></td>
</tr>
</tbody>
</table>

Cluster 1, which corresponds to 50.3% inpatient journeys, collects typical treatment behavior of unstable angina patients who have been performed PCI surgery in their treatment processes. A closer analysis shown on Table 2 indicates that cluster 1 contains typical treatment interventions (e.g., “Coronary angiography”, “Stent implantation”, etc.) for unstable angina. There is little variation occurred and common treatment events are carried out smoothly. In clinical practice, patients who are categorized into cluster 1 have shorter LOS (on average 5.79 days) than the others.

Cluster 2, which collects about 32.2% samples, contains typical conservative treatments of unstable angina. In clinical practice, patients in cluster 2 have either low risks or specific physical problems, e.g., coronary stenosis such that they prefer conservative treatments instead of PCI surgery. As a result, the average LOS of patients in cluster 2
(i.e., 8.91 days as shown in Figure 5(A)) is longer than patients in cluster 1.

Cluster 3 has captured typical treatment behaviors of unstable angina patients who have more complex conditions than others such that many treatments on the comorbidities of the patients, e.g., “Glucose regulating treatment”, “Tumor markers checks”, “Consultation”, etc., can be found in this cluster (as shown in Table 2). Note that several patients in cluster 3 are transferred to Cardiac Surgery department for the further surgical thoracotomy, such as “Coronary artery bypass graft (CABG)”, etc. Note that this variant cluster is a bit normal in the unstable angina treatment process (17.5% patients in the collections of EMRs). The average LOS of patients in cluster 3 is about 14.35 days, which is much longer than the other clusters.

4. Conclusions

In this paper, we propose a novel approach for medical inpatient journey modeling and clustering, which first collects critical treatment events in inpatient journeys from EMRs, and then develop a probabilistic model-based approach to group inpatient journeys characterized by similar treatment behaviors. To this end, we present a Bayesian HMM-based representation method, i.e., inpatient journey model, to transform the collection of medical inpatient journeys into a probabilistic space defined by the estimated posterior probabilities without losing information related to the dynamics and dependency in the data. Based on the constructed probabilistic space, similarities between inpatient journeys are calculated and the homogenous ones are grouped into the same cluster.

References
Quality Assurance of Cancer Study Common Data Elements Using A Post-Coordination Approach

Guoqian Jiang¹, Harold R. Solbrig¹, Eric Prud’hommeaux², Cui Tao³, Chunhua Weng⁴, Christopher G. Chute⁵

¹ Department of Health Sciences Research, Mayo Clinic, Rochester, MN; ² W3C/MIT, Boston, MA; ³ University of Texas Health Science Center at Houston, Houston, TX; ⁴ Columbia University, New York City, NY; ⁵ Johns Hopkins University, Baltimore, MN

Abstract

Domain-specific common data elements (CDEs) are emerging as an effective approach to standards-based clinical research data storage and retrieval. A limiting factor, however, is the lack of robust automated quality assurance (QA) tools for the CDEs in clinical study domains. The objectives of the present study are to prototype and evaluate a QA tool for the study of cancer CDEs using a post-coordination approach. The study starts by integrating the NCI caDSR CDEs and The Cancer Genome Atlas (TCGA) data dictionaries in a single Resource Description Framework (RDF) data store. We designed a compositional expression pattern based on the Data Element Concept model structure informed by ISO/IEC 11179, and developed a transformation tool that converts the pattern-based compositional expressions into the Web Ontology Language (OWL) syntax. Invoking reasoning and explanation services, we tested the system utilizing the CDEs extracted from two TCGA clinical cancer study domains. The system could automatically identify duplicate CDEs, and detect CDE modeling errors. In conclusion, compositional expressions not only enable reuse of existing ontology codes to define new domain concepts, but also provide an automated mechanism for QA of terminological annotations for CDEs.

Introduction

Domain-specific common data elements (CDEs) are emerging as an effective approach to standards-based clinical research data storage and retrieval. Notably, the National Cancer Institute (NCI) has implemented the Cancer Data Standards repository (caDSR) that adopted the ISO/IEC 11179 Metadata Registry (MDR) standard¹-². ISO/IEC 11179 defines a standard model of a meta-data registry - a registry of information about data models*. Along with the expected provenance and workflow information such as the creator, owner, workflow state, etc., part 3 of the ISO/IEC 11179 model also describes a model for formally associating data model elements with their intended meaning. It is the common meaning aspect of the 11179 standard that allows one to determine that two data elements from two different models are alternative representations of the same real world entity.

The caDSR repository has attempted to record the intended meaning of many of the data elements that support cancer study data collection and reporting which, in turn, are used to identify data elements used in relevant domains, an example of which is the Cancer Genome Atlas (TCGA) Biospecimen Core Resource (BCR) data dictionary, which is used to create clinical data collection forms for a number of clinical cancer genome study domains³. The viability and usefulness of this approach depends on the quality of the reference caDSR elements themselves. Earlier studies⁴,⁵ have uncovered serious issues with at least some of the caDSR definitions and have highlighted a need for robust, principled and automated quality assurance (QA) tools for the CDEs in cancer study domains.

Date element meanings, as recorded in the caDSR frequently use a simple form of "post-coordination", where a primary (focus) concept is identified along with one or more secondary identifiers that modify or qualify the intended target meaning. When taken on its own, this approach does not lend itself to automated validation and consistency checking. If, however, the constellation of identifiers were to be assembled into a compound expression of the sort that is used in SNOMED CT and OWL, it would then be possible use a description logic (DL) classifier to evaluate the various definitions for completeness and consistency. As an example, consider a data element about "joint pain in the hands". This could be assembled as a simple collection, with a focus concept of "pain" juxtaposed

* The 11179 standard is about metadata, not the actual data models. This is sometimes a source of confusion, as, in order to record metadata about an artifact, you also need a good idea of the type and purpose of the artifact itself. 11179 is about how one describes the provenance and purpose of a given element of a given type in a given data model and how it relates to similar elements with different types in different models.
with "joint" and "hand" or it could be constructed as a formal DL construct of "pain AND finding_site SOME hand AND finding_site SOME joint" (meaning the set of all things that simultaneously instances of (1) pain (2) have a finding_site of a hand and (3) have a finding_site of a joint). The latter form allows an automated reasoner to determine the relationship of the target data element with similar constructs. A DL reasoner would be able to determine that this data element described "hand pain" and "joint pain", was related to, but more specific than a data element that was just about "joint pain" (which could be describing ankle pain, knee pain, etc). Given sufficient information, a DL reasoner would also be able to determine that a data element about, say "joint pain in the eyebrow" was most likely nonsensical, as there was no possible instance that could simultaneously be both. While the above example may seem contrived, mistakes of this sort can and do occur in data element definitions. As an example, a data element that confuses "apoptosis" as a process with "apoptosis" as a morphological abnormality may, in turn, confuse data elements that describe genetic processes with descriptions of dead cells themselves.

DL-based mechanisms allow ontology curators to formally and unambiguously represent concept meanings and relationships, and to use off the shelf reasoning tools such as HermiT to automate the computation of the relationship between two class expressions and consistency checks. Rector, et al. developed an effective quality assurance mechanism using reasoners to incorporate qualifiers (e.g., acute or chronic) in the post-coordination in SNOMED CT. The editors of the National Cancer Institute Thesaurus (NCIt) use a DL reasoner to check the terminology completeness and consistency, and the Thesaurus is distributed as a DL based terminology. Horridge, et al. demonstrated how automated DL reasoning, along with a Justification Finding Service can be used as a QA technique for the development of large and complex ontologies such as ICD-11. The post-coordination approach using compositional expressions has been used to build a common ontology to harmonize ICD-11 and SNOMED CT. In an earlier study, we audited the semantic completeness of the SNOMED CT content using a formal model of the normal forms of SNOMED CT expressions.

The objective of the present study to design, develop and evaluate a quality evaluation tool for cancer study CDEs using a post-coordination approach. We propose a compositional expression pattern based on a modified SNOMED CT observable model, which renders the Data Element Concept model structure informed by ISO/IEC 11179. We then developed a transformation tool to convert the pattern-based compositional expressions into the OWL syntax. The OWL-based constraints are designed and the reasoning and explanation services are invoked to detect CDE modeling errors and duplicate CDEs. We then test the system utilizing CDEs from two TCGA clinical cancer study domains.

Materials and Methods

Materials

NCI caDSR CDEs, TCGA Data Dictionary and NCI Thesaurus (NCIt)

We downloaded an XML rendering of all non-retired production CDEs as of August 7, 2014 from the NCI caDSR website and an XML rendering of the snapshot of the publicly available TCGA BCR data dictionary from the TCGA website. We also downloaded the asserted (i.e. the inputs to a DL reasoner vs the conclusions) version 15.01d of NCIt in OWL format from the NCI Enterprise Vocabulary Services (EVS) website.

Semantic Web Applications and Tools

We used 4store, an open source RDF triple store as the back end for data integration and query and Protégé, an open source editor and knowledge acquisition system developed by the Stanford University in combination with HermiT version

\[\text{\textsuperscript{†}}\] "Sufficient information" in this context assumes that the reasoner would be able to process statements about negation and/or disjointness and that it its set of known facts included an axiom that explicitly or implicitly asserted that the set of things that are parts of the ear are disjoint (no pun intended) from the set of things that are joints.

\[\text{\textsuperscript{‡}}\] See: 20663007 | Apoptosis (morphological abnormality) | in the January 2014 edition of SNOMED CT.
1.3.8.3, a DL reasoner.

Methods

Figure 1 shows the architecture used in our evaluation. Module 1, Data Integration and Services combines the information from the caDSR Common Data Elements and the TCGA data dictionary as a cohesive unit, which allows the SPARQL query services to access the contents of both resources as a single unit. Module 2, Compositional Expression Transformation converts the data element meaning definitions recorded in the caDSR elements into DL expressions which become the inputs to Module 3, OWL-based Quality Assurance which uses the combination of the NCI Thesaurus and additional disjointness axioms to detect potential errors and duplications in the data element definitions. Each of these modules is described in more detail in the following sections.

Module 1: Data Integration and Services

The caDSR CDEs and TCGA data dictionary were converted from XML into RDF using the XML2RDF tool that was developed for the Redefer project. The output was loaded into the 4Store RDF triple store instance and a SPARQL endpoint was added to enable the actual SPARQL query services. The original XML files and transformed RDF files can be found at the project github website (https://github.com/caCDE-QA/owl-qa).

Figure 2 shows a SPARQL query that retrieves all CDEs from the domain “clinical pharmaceutical” and their data related to data element concept, object class and property recorded in caDSR.

Module 2: Compositional Expression Transformation

ISO/IEC 11179 identifies a basic model of a definition of a Data Element Concept, "A concept that is an association of a property ("a quality common to all members of an object class") with an object class ("set of ideas, abstractions or things in the real world that are identified with explicit boundaries and meaning and whose properties follow the same rules")². The data element definitions in the NCI caDSR separated the property and object class concept references, an aspect that we were able to take advantage of in an earlier study. One of the defining concepts in both the property and object class definitions were identified as the "focus concept" (i.e. one focus property and one focus object class). The relationship between the remaining concepts and the focus was left unspecified. Figure 3 shows the 11179 Data Element Concept model region and Table 1 shows two sample Data Element Concepts as defined in the NCI caDSR, with the primary or focus concept of each definition in bold. The first definition asserts that the object of the data element was Internal Radiation Therapy and the property being measured was the "Technique" qualified by "delivery". In the second example, the object is asserted to be "blood pressure", qualified by "person" and the property of the blood pressure was "assessment".

It became apparent that we were dealing with two levels of modeling. On the most fundamental level, we were working with data elements as defined in ISO 11179 ("unit of data that is considered in context to be indivisible")². In this context, the data elements "Adjuvant Postoperative Pharmaceutical Therapy Administered Indicator" (3397567), "Year of Death" (2897030) and "Initial Pathology Diagnosis Method" (2757948) all have the

[Diagram of Data Element Concept Definition]
same object class/property definitional structure. Considerable value could be added, however, if the data elements could be correctly positioned in more sophisticated model such as the proposed SNOMED CT observables model\textsuperscript{18} as shown in Figure 4, which would allow the various property and object class modifiers to be correctly identified by the role they played in context of the data element itself. One of the challenges in doing this, however, is that not all of the data elements in the caDSR can be treated as observation results. As an example, "Tumor Tissue Site" (CDE 3427536) is not an observation unto itself but would instead serve as a direct site or (inherent) location in a more complete location.

Table 1. Sample Data Element Concept Definitions

<table>
<thead>
<tr>
<th>Data Element Id</th>
<th>Name</th>
<th>Object Class Concepts</th>
<th>Object Class Concept Meanings</th>
<th>Property Concepts</th>
<th>Property Concept Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2201422</td>
<td>Brachytherapy Delivery Technique</td>
<td>C15195</td>
<td>Internal Radiation Therapy</td>
<td>C16847 C61560</td>
<td>Technique Delivery</td>
</tr>
<tr>
<td>2004291</td>
<td>Diastolic Blood Pressure</td>
<td>C54706 C25190</td>
<td>Blood Pressure Person</td>
<td>C25367</td>
<td>Assessment</td>
</tr>
</tbody>
</table>

The remainder of this paper focuses on the first and more abstract model level, where every CDE is treated as complete "observation" unto itself -- either as a property of an object (independent continuant) or as characterizing a process.

Figure 5 shows a diagram illustrating the mappings between a modified version of the SNOMED CT observable model and the ISO/IEC 11179 model implemented in the NCI caDSR. As illustrated in the figure, a data element concept is mapped with an observable entity; a data element concept property is mapped to the target (i.e., range class) of the predicate “is about”; a primary property is mapped to the target (i.e., range class) of the predicate “property type”; a data element object class is mapped to the target of the predicate “inheres in or characterizes”. Note that we combined two predicates “inheres in” and “characterizes” into a single predicate “inheres in or characterizes” as NCI caDSR does not directly provide the distinction between independent continuants and processes (see the section Discussions for more details). Within the NCI caDSR model, a primary object class is the target of the predicate “object class type” and a primary property is the target of the predicate “property type”.

We created a transformation tool using the Jena Java API\textsuperscript{19} that takes the input from the results of a SPARQL query described in the previous section and renders the data recorded for the data element concept of a CDE into an OWL-based compositional expression. Table 2 shows two compositional expressional examples transformed to represent the semantics for the data element concept, object class and property of the CDEs “Clinical Trial Drug Classification Name” and “Agent Administration Total Dose Code”.

Figure 4. Proposed SNOMED CT Observation Result Model

Figure 5. Mappings between SNOMED CT Observable Model and ISO/IEC 11179
Module 3: OWL-based Quality Assurance

Module 2 transforms the CDE definitions into OWL compositional expressions. We are now in a position that we can use a DL reasoner to validate the expressions against the underlying ontology.

Table 2. Compositional Expression in caDSR and enhanced OWL representation in Manchester OWL Syntax

<table>
<thead>
<tr>
<th>Original Data Recorded in caDSR</th>
<th>Transformed Compositional Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Id: 3378323</td>
<td>Class: ‘Clinical Trial Drug Classification Name’</td>
</tr>
<tr>
<td>CDE Name: Clinical Trial Drug Classification Name</td>
<td>Annotations:</td>
</tr>
<tr>
<td>Property Code: C25161</td>
<td>label &quot;Clinical Trial Drug Classification Name&quot;</td>
</tr>
<tr>
<td>Property Name: Classification</td>
<td>EquivalentTo:</td>
</tr>
<tr>
<td>Primary Property: C25161</td>
<td>'Observable Entity'</td>
</tr>
<tr>
<td>Object Class code: C71104:C1708</td>
<td>and ('is about' some (Classification and ('inheres in or characterizes' some ('Clinical Trial Agent'</td>
</tr>
<tr>
<td>Object Class Name: Clinical Trial Agent</td>
<td>and ('object class type' some Agent)) and ('property type' some Classification)))</td>
</tr>
<tr>
<td>Primary Object Class: C1708</td>
<td>SubClassOf: 'Observable Entity'</td>
</tr>
<tr>
<td>Public Id: 3088785</td>
<td>&quot;Class: 'Agent Administration Total Dose Code'</td>
</tr>
<tr>
<td>CDE Name: Agent Administration Total Dose Code</td>
<td>Annotations:</td>
</tr>
<tr>
<td>Property Code: C25304:C25488:C25709:C25162</td>
<td>label &quot;Agent Administration Total Dose Code&quot;</td>
</tr>
<tr>
<td>Property Name: Total Dose Unit of Measure Code</td>
<td>EquivalentTo:</td>
</tr>
<tr>
<td>Primary Property Name: C25162</td>
<td>'Observable Entity'</td>
</tr>
<tr>
<td>Object Class code: C70962</td>
<td>and ('is about' some (Quantity and ('inheres in or characterizes' some ('Prescription Agent' and ('object class type' some Agent)) and ('property type' some Quantity)))</td>
</tr>
<tr>
<td>Object Class Name: Agent Administration</td>
<td>SubClassOf: 'Observable Entity'</td>
</tr>
<tr>
<td>Primary Object Class: C70962</td>
<td></td>
</tr>
</tbody>
</table>

Checking for duplicate data elements

A DL reasoner can find equivalent (duplicate) definitions. This provides an automated mechanism for checking potential duplicate data elements that have the same meanings.

Mechanism for detecting data element modeling errors

With the OWL rendering of compositional expression patterns based on the SNOMED CT observable model, we can add constraints to the patterns that will allow a DL reasoner to detect potential data element modeling errors. The first constraint is to assert that instances of Object Class and Property are disjoint -- that nothing can simultaneously be a thing or process and attribute or characteristic of the thing or process.

A second constraint is an assertion about the subjects (domain) and objects (range) of the :is_about, :property_type, :measures, and :object_class_type predicates. The domain/range constraint restricts each object property linking between the instances of asserted classes. Table 3 shows two types of the OWL constraints asserted in the compositional patterns.

Explanation of detected inconsistencies or errors

We used the Protégé 5.0 beta version and the HermiT reasoner which can explain the possible reasons for inconsistencies or errors. Figure 6 shows a screenshot of illustrating the explanations for the class “C1708:Agent” violating the disjointness constrain asserted between the classes “Observable Entity Object Class” and “Observable Entity Property”.

Figure 6. Example Explanation of an Inconsistency
Table 3. Two types of the OWL constraints asserted in the compositional patterns in Manchester OWL syntax

<table>
<thead>
<tr>
<th>disjointness constraints</th>
<th>domain/range constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class: 'Observable Entity Property'</td>
<td>ObjectProperty: 'is about'</td>
</tr>
<tr>
<td>DisjointWith: 'Observable Entity Object Class', 'Primary Object Class'.</td>
<td>Domain: 'Observable Entity'</td>
</tr>
<tr>
<td>Class: Primary Object Class'</td>
<td>Range: 'Observable Entity Property'</td>
</tr>
<tr>
<td>DisjointWith: 'Observable Entity Property', 'Primary Property'.</td>
<td>ObjectProperty: 'property type'</td>
</tr>
<tr>
<td>Class: 'Primary Property'</td>
<td>Domain: 'Observable Entity Property'</td>
</tr>
<tr>
<td>DisjointWith: 'Observable Entity Object Class', 'Primary Object Class'.</td>
<td>Range: 'Primary Property'</td>
</tr>
</tbody>
</table>

Case study of CDEs from two TCGA cancer genome study domains

We used the configuration above perform a quality evaluation of the CDEs from two clinical cancer domains in TCGA data dictionary: Clinical Pharmaceutical and Clinical Shared. CDEs from each domain along with the data element concept, object class and property assertions were retrieved using the SPARQL query in Figure 2. The results were converted compositional expressions and classified using the Protégé based DL reasoner. We were able to identify the equivalent CDEs and the CDEs violating the reasoning constraints. To verify the modeling errors, four co-authors (GJ, HS, CT, CW) reviewed the detected CDEs by checking the definitions of the CDEs recorded in the NCI caDSR to verify whether there is a modeling error or not.

Results

In total, TCGA data dictionary contains 775 CDEs for 38 clinical cancer domains, which cover 21 cancer types. In the present study, we performed a case study of two clinical cancer domains: Clinical Pharmaceutical and Clinical Shared, which contain 18 and 98 CDEs respectively.

The reasoning services identified 6 CDEs with equivalent CDEs from the domain Clinical Pharmaceutical and 29 CDEs with equivalent CDEs from the domain Clinical Shared. In total, there are 12 groups of equivalent CDEs. Human-based review shows that among 12 groups of equivalent CDEs identified, the CDEs in 2 groups had modeling errors, indicating that they should not be considered as the equivalent CDEs. Table 4 shows the equivalent CDE groups identified with errors from the domain Clinical Shared.

Table 4. The equivalent CDEs identified with errors from the domain Clinical Shared

<table>
<thead>
<tr>
<th>Domain</th>
<th>Equivalent CDEs</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Shared</td>
<td>2181650 Patient Smoking History Category 2228604 Started Smoking Year</td>
<td>Yes</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>88 Performance Status Assessment Eastern Cooperative Oncology Group Scale 2792763 Performance Status Assessment Timepoint Category 2003853 Karnofsky Performance Status Score</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 5 shows the results of CDEs violating constraints from two TCGA domains by reasoning services. In total, there are 19 CDEs (out of 116 CDEs) identified with constraint violations. Human-based review shows that all 19 CDEs had modeling errors in their asserted primary properties. Table 5 provided the suggested primary properties for the 19 CDEs.

Discussion

In a previous study, we analyzed the terminological concepts associated with the standard structure of the caDSR CDEs using the UMLS Semantic Network. We found out that the semantic annotations of a CDE that did not observe the overall pattern of disjointness between dominant semantic types of primary object class/property, had a high probability to have modeling errors. Although the ISO/IEC 11179 specification states that it provides a semantically precise structure for data elements, the standard does not specify disjointness constraints between
object class concept and property concept. In the present study, we designed a compositional expression pattern to post-coordinate the data element concept of a CDE and transformed them to the Semantic Web OWL-based expression. The transformation allows us to take advantage of the built-in feature of OWL in expressing the disjointness and domain/range restrictions among a set of OWL classes, and subsequently invoke the reasoning services provided in existing OWL-DL reasoning tools (e.g., the HermiT reasoner used in this study) to check the inconsistencies and violations. This approach has proven to be very useful in identifying potential modeling errors. With the case study of the CDEs from two TCGA clinical cancer study domains, we reviewed those CDEs violating the constraints (n=19) and found that all of them had incorrect property concepts asserted.

Table 5. CDEs violating constraints identified from two TCGA domains by reasoning services.

<table>
<thead>
<tr>
<th>Domain</th>
<th>CDEs Violating Constraints</th>
<th>Asserted Primary Property</th>
<th>Semantic Type</th>
<th>Our Suggested Primary Property</th>
<th>Semantic Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Pharmaceutical</td>
<td>2975232 Prior Therapy Regimen Text</td>
<td>C1708/Agent</td>
<td>Chemical Viewed Functionally</td>
<td>C25365/Description</td>
<td>Intellectual Product</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2791194 First Disease Recurrence Disease Extent Category</td>
<td>C13717/Anatomic Site</td>
<td>Body Location or Region</td>
<td>C25372/Category</td>
<td>Classification</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>3108203 Neoplasm Anatomic Subdivision Name</td>
<td>C13717/Anatomic Site</td>
<td>Body Location or Region</td>
<td>C42614/Name</td>
<td>Conceptual Entity</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>3124503 First Recurrent Non-Nodal Metastatic Anatomic Site Descriptive Text</td>
<td>C13717/Anatomic Site</td>
<td>Body Location or Region</td>
<td>C25365/Description</td>
<td>Intellectual Product</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>3427536 Tumor Disease Anatomic Site</td>
<td>C13717/Anatomic Site</td>
<td>Body Location or Region</td>
<td>C25365/Description</td>
<td>Intellectual Product</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2006657 Diagnosis Age</td>
<td>C15220/Diagnosis</td>
<td>Diagnostic Procedure</td>
<td>C2515/Age</td>
<td>Organism Attribute</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2896956 Month Cancer Initial Diagnosis Number</td>
<td>C15220/Diagnosis</td>
<td>Diagnostic Procedure</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2896958 Day Cancer Initial Diagnosis Number</td>
<td>C15220/Diagnosis</td>
<td>Diagnostic Procedure</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2896960 Year Cancer Initial Diagnosis Number</td>
<td>C15220/Diagnosis</td>
<td>Diagnostic Procedure</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2896991 Month Tumor Recurrence After Initial Treatment Number</td>
<td>C15220/Diagnosis</td>
<td>Diagnostic Procedure</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2897006 Day Tumor Recurrence After Initial Treatment Number</td>
<td>C15220/Diagnosis</td>
<td>Diagnostic Procedure</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2897008 Year Tumor Recurrence After Initial Treatment Number</td>
<td>C15220/Diagnosis</td>
<td>Diagnostic Procedure</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>3382736 Prior Cancer Diagnosis Occurrence Description Text</td>
<td>C15220/Diagnosis</td>
<td>Diagnostic Procedure</td>
<td>C25365/Description</td>
<td>Intellectual Product</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2897014 Month Tumor Progression After Initial Treatment Number</td>
<td>C15368/Treatment</td>
<td>Health Care Activity</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2897016 Day Tumor Progression After Initial Treatment Number</td>
<td>C15368/Treatment</td>
<td>Health Care Activity</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2897018 Year Tumor Progression After Initial Treatment Number</td>
<td>C15368/Treatment</td>
<td>Health Care Activity</td>
<td>C25164/Date</td>
<td>Temporal Concept</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2178045 Age Began Smoking in Years</td>
<td>C18270/Cigarette Smoking</td>
<td>Individual Behavior</td>
<td>C2515/Age</td>
<td>Organism Attribute</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2554674 Patient Death Reason</td>
<td>C28554/Death</td>
<td>Finding</td>
<td>C25365/Description</td>
<td>Intellectual Product</td>
</tr>
<tr>
<td>Clinical Shared</td>
<td>2584114 Primary Other Site of Disease Name</td>
<td>C2991/Diseases and Disorders</td>
<td>Disease or Syndrome</td>
<td>C25365/Description</td>
<td>Intellectual Product</td>
</tr>
</tbody>
</table>

While the domain expertise would be definitely helpful in evaluating whether an object class/property concept is correctly asserted or not, the challenging question is how to make the system to tell automatically that a terminological code could only be used for annotating an object class not a property. We consider that the approaches leveraging the upper level ontologies such as UMLS Semantic Network, basic formal ontology (BFO)\(^{20, 21}\) or BioTop ontology (a top-domain ontology for the life science)\(^{22, 23}\) would potentially provide a formal approach to define the constraints for supporting the CDE modeling applications. Linking to ISO/IEC 11179 model, a
constraint can be made like “an object class concept has to be an independent continuant or processual entity whereas a property concept cannot be such entity”. Actually, the SNOMED observable model we referenced in this study has been developed using such principles. As we are using a modified version of SNOMED observable model to post-coordinate the data element concepts, we plan to look into the upper level ontology-based approaches further in our future study.

With the compositional expressions coded for the data element concepts representing the meaning of CDEs, the Semantic Web reasoning services can be invoked to automatically identify semantically equivalent CDEs. On the one hand, this powerful feature could be used to detect potential duplicate CDEs as well as the modeling errors. As demonstrated in our case study (in Table 4), among 12 groups of equivalent CDEs are identified, the CDEs in 2 groups had modeling errors, meaning that they should not be considered as the equivalent CDEs. For example, reviewing the definition for the CDEs “2181650|Patient Smoking History Category” and “2228604|Started Smoking Year”, their primary property concept are all assigned as the “Patient Medical History” which is not correct. If we correct the primary property concepts for the two CDEs as “C25372|Category” and “C25164|Date” respectively, they will be no longer inferred as the equivalent CDEs. On the other hand, the post-coordination approach could be used to harmonize and standardize the data elements collected from different study groups or sites. Specifically, if we could use the compositional expression pattern to post-coordinate these data elements using standard ontologies such as NCIt, we would be able to accurately identify the data elements with the same meanings and easily harmonize the data element and enable data integration. This is exactly the very goal the ISO/IEC 11179 standard would like to achieve. Leveraging the SNOMED CT observable model would help make the compositional expression pattern user-friendlier. For example, the notion of Observable Entity is user-friendlier to clinical study researchers than the notion of “Data Element Concept”.

We have demonstrated the value of using Semantic Web technologies in building our QA tools for cancer study CDEs. First, we developed a semantic metadata repository using an open source RDF triple store. The RDF-based data model not only provides a scalable and powerful framework for data integration, but also provides standard SPARQL query services that allow us to easily retrieve the preferred set of CDEs in a particular TCGA clinical cancer study domain. In a separate study, we demonstrated that such Semantic Web-based metadata repository could also be used to support authoring detailed clinical models (DCMs) in clinical cancer study domains 24. In the present study, the framework enables us to build a domain-specific QA mechanism. Although we focused on the two TCGA domains as a case study, the approach could be easily generalized to any other TCGA domains. Second, we developed a compositional expression transformation tool that transforms the data structure of a data element concept informed by ISO/IEC 11179 into an OWL-based representation. The compositional expression is based on a modified version of SNOMED CT observable model. We consider that the syntax of SNOMED CT’s compositional expression grammar 25 may be useful as an intermediate layer in our transformation tool, which will provide a human readable format and potentially improve the usability of authoring a post-coordination expression. Third, as the NCI Thesaurus is distributed in the OWL format, we can easily integrate the compositional expressions generated for the CDEs in each domain with the NCIt using the “owl:imports” mechanism. As the latest version of the NCIt has marked all its retired concepts using the deprecated annotation “owl:deprecated”, we were able to detect retired concepts out of the compositional expressions that are originally based on the annotations using the NCIt codes. Fourth, Semantic Web reasoning services are a critical component of our QA tool. We used the latest version of HermiT reasoning plugin in Protége 5 environment in this study. We found that incremental reasoning and explanation services are very helpful in detecting and explaining the CDE modeling errors and duplicate CDEs. However, further studies will be needed in transforming the reasoning results into the well-formed report to inform the decision-making of CDE curators. In addition, integration of compositional expressions with the large-scale ontology such as NCIt will make the reasoning services more complicated and also require high performance. There already have a number of high-performance reasoners such as Snorocket 26 available in the Semantic Web community, and we will compare and test out such reasoners in the future.

There are a number of limitations in the study. First, our QA tool is based on the CDEs from a particular domain. While the domain-specific approach has advantage of making the auditing results more interpretable, some of the modeling errors may not be able to be detected. The systematic solutions would include 1) enabling the violation checking globally on CDEs across domains; 2) implementing the upper level ontology-based approach as discussed above. Second, we used a simplified compositional expression model that captures main constructs of a data element concept informed by ISO/IEC 11179 standard. For example, for an object class concept “Adjuvant Hormone Therapy”, it is recorded as “C2140:C15445”. The primary object class “C15445|Endocrine Therapy” has been captured in our model but the semantic relationship between “C2140|Adjuvant” has not been captured. We plan to
look into the approach for post-coordination such expressions in the future. Third, the QA of the data structure describing the meaning of a value domain is out of scope in the present paper and will be conducted in a separate study.

Conclusion

In this study, we developed and evaluated a QA tool for cancer study CDEs using a post coordination approach. We designed a compositional expression pattern based on a version of SNOMED CT observable model, which is used to represent the data structure of a data element concept (i.e., the meaning of data element) informed by the ISO/IEC 11179 metadata standard. Leveraging the existing Semantic Web tools, we demonstrated that the post-coordination approach could enable an effective and automated mechanism in detecting potential CDE modeling errors and duplicate CDEs. Future work will be focused on 1) developing a systematic QA approach leveraging upper level ontologies; 2) refining the compositional expression model by aligning with SNOMED observable model; 3) making the reasoning-based explanation more user-friendly; and 4) incorporating the value domain in the scope.

Acknowledgements

The study is supported in part by a NCI U01 Project – caCDE-QA (1U01CA180940-01A1).

References

15. 4store. 2015 [February 20, 2015]; Available from: http://4store.org/.
In Search of Social Translucence: An Audit Log Analysis of Handoff Documentation Views and Updates

Silis Y. Jiang, MA¹, R. Stanley Hum, MD², David Vawdrey, Ph.D.¹,³, Lena Mamykina, Ph.D.¹
¹ Department of Biomedical Informatics, Columbia University, New York, NY
² Department of Pediatrics, Columbia University, New York, NY
³ New York Presbyterian Hospital Value Institute, New York, NY

Abstract

Communication and information sharing are critical parts of teamwork in the hospital; however, achieving open and fluid communication can be challenging. Finding specific patient information within documentation can be difficult. Recent studies on handoff documentation tools show that resident handoff notes are increasingly used as an alternative information source by non-physician clinicians. Previous findings also show that residents have become aware of this unintended use. This study investigated the alignment of resident note updating patterns and team note viewing patterns based on usage log data of handoff notes. Qualitative interviews with clinicians were used to triangulate findings based on the log analysis. The study found that notes that were frequently updated were viewed significantly more frequently than notes updated less often \(p < 2.2 \times 10^{-16}\). Almost 44% of all notes had aligned frequency of views and updates. The considerable percentage (56%) of mismatched note utilization suggests an opportunity for improvement.

Introduction

Care teams require communication and information sharing to coordinate and provide care for patients in the hospital¹⁻³. However, poor communication is frequently associated with errors and sentinel events⁴. Studies on improving patient outcomes and care processes after the implementation of electronic health records (EHRs) have shown both positive and negative results⁵⁻⁸. Critical patient information is often difficult to obtain quickly or is simply missing in the EHR⁹,¹⁰. Moreover, team member information needs differ according to clinician training and specialization¹¹. Clinical documentation, a common form of team communication, often contains out of date information or is simply not read¹²,¹³. These and other problems pose a challenge to adapting EHR based tools to facilitate team-based communication and information sharing.

There is, however, emerging evidence that new electronic tools for facilitating handoffs, handoff documentation tools, are increasingly used as a source of patient information¹⁰,¹³,¹⁴. Schuster et al. found that despite handoff notes being written by and for residents, other services, such as ancillary services, also consulted these notes for information¹⁰. Furthermore, Vawdrey et al. demonstrated that despite its original intent to facilitate transitions of care for resident physicians, non-physician services, such as nursing and pharmacy, make up more than half of all views of handoff notes¹³.

In a follow-up study to Vawdrey et al., residents, the primary authors of handoff notes, acknowledged team members reading their notes¹⁴. Some even acknowledged a change in the frequency of updates to their handoff notes based on how others in the team incorporated handoff notes into their daily workflow¹⁴. However, many questions still remain about using handoff documentation as a way to disseminate information to other care team members. One such question is to what degree are residents’ expectations and practices in regards to updating handoff notes aligned with other team members? In this study, we examine the patterns of updates and views for electronic handoff notes (handoff notes from this point forward) at a large tertiary care hospital. Specifically, this study investigates the degree with which users are aligned in their viewing and updating practices and focuses on the following question: Are more frequently updated notes also viewed more frequently?

Materials and Methods

Handoff Tool

Handoff Tool is a custom-designed module included in the commercial EHR system (Sunrise Eclypsis). At the time the study data was collected, Handoff Tool included 9 free text boxes: Active Issues, Consult Notes, Contact Information, Coverage To Do List, Discharge Planning, Hospital Course, Notes and Comments, Patient Summary, and Primary To Do List. The tool also provided the functionality to produce a printable report that included structured data, such as labs and medication. Furthermore, users could choose to print a cover sheet that summarized
selected patients. At the time of the study, only physicians, physician assistants, and nurse practitioners could make edits to the handoff document, but all members of the care team could access and view the document. In contrast to traditional notes, such as the daily progress notes, which started blank each shift, handoff notes were updated from the last save iteration of the note. The tool is described in further detail in[1].

Dataset

To examine updating and viewing practices related to handoff notes, we collected the document-updating and document-viewing event logs for all of October 2013 at Columbia University Medical Center campuses of NewYork-Presbyterian Hospital, a large urban, academic medical center. The document-editing event log contained the date and time of each edit event, along with which free-text box was edited, the patient medical record number, the patient bed location, the clinician user ID, and the care provider role. The document-viewing event log contained the date, time, and duration of each handoff document access. Additional information in the log included the patient medical record number, the patient bed location, the clinician user ID, and the care provider role. The dataset was then divided into individual notes for each patient. Each unique handoff document was identified by the patient medical record, and each shift was defined as being either from 7am to 7pm (day shift) or 7pm to 7am the next day (night shift). For each document, the number of updates and views were tabulated. An update was defined as an event when the handoff document was edited. A view was defined as instance when a clinician, other than the primary author of the note, opened and closed the note. The primary author for a note was defined as the clinician who made the most edits to the handoff note during a particular shift. Furthermore, the frequency of views for each clinician role for each note during each shift was also calculated. To correlate the frequency of views and updates, each document was assigned an activity level score, which reflects the relative frequency of the views and updates for each document (see Table 1).

Table 1. Activity Level Coding Schema

<table>
<thead>
<tr>
<th>Updates/Views</th>
<th>None (0 Updates)</th>
<th>Minimal (1-2 Updates)</th>
<th>Frequent (3 or More Updates)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (0 views)</td>
<td>A (matched case)</td>
<td>D (more updates)</td>
<td>E (more updates)</td>
</tr>
<tr>
<td>Minimal (1-2 views)</td>
<td>G (more views)</td>
<td>B (matched case)</td>
<td>F (more updates)</td>
</tr>
<tr>
<td>Frequent (3 or more views)</td>
<td>H (more views)</td>
<td>I (more views)</td>
<td>C (matched case)</td>
</tr>
</tbody>
</table>

Quantitative Analysis

To identify temporal patterns of viewing, the hour of each view was extracted and plotted. To identify the relationship between updates and views, a linear regression was fitted on the frequency of views and updates for each patient’s note per shift. Then each note was stratified into one of three categories based on the frequency of updates according to the schema found in Table 1. An ANOVA (and Kruskal-Wallis) test was then performed to identify if there was a difference in frequency of note views between the three levels of note updates.

The statistical packages pandas, numpy, and scipy for Python were used to analyze the data, and ggplot2 for R was used to visualize the data.

Qualitative Analysis

To triangulate the finding from the quantitative analysis, interviews were conducted, transcribed, and analyzed for common themes. In total, 3 interviews were conducted. Two residents from the adult general medicine department, and one nurse practitioner from a pediatric hematology/oncology unit were interviewed. All clinicians were consented prior to interview. This study was approved by the Columbia University Medical Center Institutional Review Board (CUMC IRB).

Results

Description of Handoff Note Views

In total, 70,042 handoff notes were included in the analysis. Handoff notes were viewed an average of 2.34 times per shift. When aggregated among all notes, physicians viewed handoff notes most frequently (207,849 times), followed
by physician assistants (46,792) and nurses (40,593). A complete breakdown of the frequency of views by common clinical roles can be found in Figure 1. The data shows that while physicians account for the majority of note views, all other roles combined account for 36.41% of views. During the interviews, all participants explained that their handoff notes were written primarily for clinicians who are covering during the evening shift. However, all three participants also acknowledged that others, such as consulting physicians or bedside nurses also read the note.

![Note Views Among Clinical Roles](image)

**Figure 1.** Monthly Handoff Note Views Among Common Clinical Roles.

Temporally, 38.70% of handoff views occurred during handoff periods (6am to 8am and 5pm to 7pm) as shown in Figure 2. This shows that while handoff notes are most frequently viewed during handoff periods, most views do not actually occur during time periods commonly associated with handoff.

**Relationship Between Views and Updates**

To analyze the data using linear regression and ANOVA, three outlier data points (notes where updates were greater than 50 during a shift), were removed. As shown in Figure 3A, notes with more updates were also viewed more frequently. In Figure 3A, the $R^2$ value is 0.2611 with a p-value $< 2.2 \times 10^{-16}$. The data was then stratified into three groups by updates (no updates, minimal updates, and frequent updates). The ANOVA test showed that notes in the different groups had significantly different frequency of views ($F = 12893$, $p < 2.2 \times 10^{-16}$). Post-hypothesis testing using Tukey’s HSD test reveals that all three groups are significantly different from each other ($p < 2.2 \times 10^{-16}$ for each instance). The greatest mean difference between groups is between frequently updated notes and notes not updated during that shift (difference = 3.39 views per shift). The mean difference in note views between notes that are minimally updated and frequently updated is also similarly large.
Given that notes in the frequently updated group were viewed more frequently than notes in the other two groups, those notes were analyzed further. In this analysis, the focus shifted from the frequency of note views to the frequency of updates. Table 2 shows the frequencies of note updates for different units of the hospital. The left hand
side of the table shows the top 15 units that contributed the most frequently updated handoff notes. The table shows that frequently updated notes come from many different units. The right hand side of the table shows the top 15 units where a high proportion of the handoff notes were frequently updated. Notably, several of these units are intensive care units (B09N, B09S, MICA, M8GS, M4SI, AICU) that provide care for complex and unstable patients. This finding indicates that use of handoff notes as part of the clinical workflow may be at least partially dependent on unit practices and norms and patient severity.

Table 2. Unit Characteristics of Frequently Updated Notes

<table>
<thead>
<tr>
<th>Top Units Contributing Frequently Updated Notes</th>
<th>Top Units with Notes that are Frequently Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit</td>
<td>Service(s)</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>B06T</td>
<td>Pediatric Cardiology/Neurology/Pulmonolgy</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>B04T</td>
<td>Pediatric Medical/Surgical/GI/Transplant</td>
</tr>
<tr>
<td>M6GS</td>
<td>Adult Medicine/Cardiology/Neurology</td>
</tr>
<tr>
<td>M8GS</td>
<td>Adult Neurologic Intensive Care Unit</td>
</tr>
<tr>
<td>M5CC</td>
<td>Adult Cardiac Care Unit</td>
</tr>
<tr>
<td>B09T</td>
<td>Pediatric Cardiac Intensive Care Unit</td>
</tr>
<tr>
<td>M7GS</td>
<td>Adult Cardiology</td>
</tr>
<tr>
<td>B05T</td>
<td>Pediatric Hematology/Bone Marrow Transplant</td>
</tr>
<tr>
<td>M5GN</td>
<td>Adult Medicine/Cardiology/Neurology</td>
</tr>
<tr>
<td>M5CT</td>
<td>Adult Cardiothoracic Intensive Care Unit</td>
</tr>
<tr>
<td>M6GN</td>
<td>Adult Medicine</td>
</tr>
<tr>
<td>M9GS</td>
<td>Adult Medicine</td>
</tr>
<tr>
<td>B09S</td>
<td>Pediatric Intensive Care Unit</td>
</tr>
<tr>
<td>M6HN</td>
<td>Adult Medicine</td>
</tr>
<tr>
<td>M6GS</td>
<td>Adult Medicine</td>
</tr>
</tbody>
</table>

**Bolded** units overlap on both side of Table 2.

**Matched Utilization of Handoff Notes**

To further analyze the use of Handoff Tool from the team perspective, the relative note activity level was coded according to the schema in Table 1. Tables 3 and 4 show that matched cases of handoff views and updates are the most prevalent (43.44%). Within matched utilization, notes were most likely to be either not used in a shift at all (0
updates and views) or used frequently during a shift (Table 3). Between the two types of mismatched utilization, notes were more likely to have relatively more views than updates during a shift (Table 4).

Table 3. Percentage of Notes By Note Activity Level

<table>
<thead>
<tr>
<th>Updates/Views</th>
<th>None</th>
<th>Minimal</th>
<th>Frequent</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>18.10%</td>
<td>12.19%</td>
<td>4.10%</td>
</tr>
<tr>
<td>Minimal</td>
<td>19.60%</td>
<td>8.73%</td>
<td>6.98%</td>
</tr>
<tr>
<td>Frequent</td>
<td>7.08%</td>
<td>6.50%</td>
<td>16.71%</td>
</tr>
</tbody>
</table>

Table 4. Percentage of Notes By Match/Mismatch Category

<table>
<thead>
<tr>
<th>Matched</th>
<th>Mismatched (More Updates)</th>
<th>Mismatched (More Views)</th>
</tr>
</thead>
<tbody>
<tr>
<td>43.44%</td>
<td>23.20%</td>
<td>33.10%</td>
</tr>
</tbody>
</table>

Qualitative Analysis

The qualitative interviews provided much insight into log data findings. All three interview participants explained that they primarily wrote handoff notes to provide information for night coverage teams and consulting physicians. Participant 3 expressed that bedside nurses frequently read the participant’s handoff notes at the beginning of the shift and often asked her follow-up questions based on the handoff note. This is consistent with qualitative findings from previous work by Jiang et al. showing that intensive care unit residents are aware that other team members, such as bedside nurses, are reading their handoff notes. Participant 2 explained that, “I think that if there had been a greater awareness of others reading [the handoff note], it would have been updated more.” This suggests that awareness of other’s actions acts as a positive feedback loop in regards to updating. However, participants 1 and 2 described little to no interactions with other clinicians prompted by the handoff note, suggesting that unit practice variations impacted the use of handoff note for information sharing. Expounding further upon the differences in units, participant 2 explained that ICU teams and general medicine teams operated differently. ICU teams often operated closely, while general medicine teams operated in a more diffuse manner. This difference may also provide one explanation for why frequently updated notes are also viewed more often.

Discussion

Previous studies have suggested that nurses and ancillary services used resident handoff notes as information sources. Further pursuing this line of investigation, this study investigated how handoff notes are used from a team perspective, and on their possible role in facilitating team communication. To complement previous qualitative studies on handoff notes, this study took a quantitative focus. The findings showed that handoff notes are read throughout the day by multiple care team members. Most of these note viewers were physicians or physician assistants; however, nurses also made up a noticeable number of the user base. As highlighted by the interviews, the frequency of notes updates is at least partially due to the awareness that other team members, such as consulting physicians or nurses are referring to them for information.

Currently, many clinicians rely on verbal communication in transmitting or obtaining patient information because documentation is not updated sufficiently quickly. This study found that units with the almost critical patients tend to have notes that are frequently updated and viewed. A possible interpretation of this finding is that those units are most likely to include Handoff Tool as one method to disseminate patient information to the rest of the team. This conclusion is mirrored in the interviews, which revealed that teams in ICUs often rely on Handoff Tool for documenting and sharing information more than general medicine units. This finding suggests a possible shift towards greater utilization of documentation, particularly electronic handoff notes, to pass on patient information.

Overall the study revealed that there exists a high degree of alignment in editing/viewing practices for electronic handoff notes. This is particularly the case for intensive care units, where patients’ conditions often change rapidly throughout a shift, and where members of patient care teams must work in close alignment with each other. Previously, researchers used the term social translucence to describe the degree of awareness between individuals about their respective practices in social settings. According to Kellogg et al, being aware of intentions and actions
of others helps individuals to maintain a high degree of cohesion and coordination. This study suggests that across the hospital, and particularly for intensive care units, electronic handoff tools are contributing to social translucence among members of patient care teams. Residents update their notes more frequently knowing that others refer to these notes for updated patient information, and other members of the teams view these notes more frequently with the expectation that they are updated. Handoff notes at our institution are carried over from the previous version and act more similarly to a single note that evolves over time; rather than traditional notes which are more similar to a collection of individual notes about a patient over time. Because of this difference, alignment of note update and view frequencies is an important indicator that note writers are providing and note viewers are reviewing updated information. However, the study also showed that there are cases when these practices are not aligned, suggesting an opportunity for improvement. For example, currently, clinicians are not informed of new updates to handoff notes that might indicate important changes in patients’ conditions. One possible solution to this limitation is the addition of an RSS-style subscription of handoff note updates that could help clinicians more reliably disseminate new patient information and maintain higher level of awareness within patient care teams.

Limitations
A key limitation to the study is that it only captured views of the electronic version of the handoff note, but not the views of its printed version. During the interviews clinicians repeatedly expressed that they frequently consulted the printed version of the handoff note throughout the day in lieu of the electronic version. Another limitation in the study is the small sample of participants in the interviews. However, these interviews were primarily conducted to help interpret the quantitative findings, rather than to draw novel conclusions. Lastly, this dataset represents a small temporal period within a single academic medical center. Handoff practices constantly evolve, and not all variations of these practices were captured with the current dataset.

Conclusion
This study examined and compared patterns of updates and views of electronic handoff notes based on the usage log data. The study revealed that use of handoff notes was commonly aligned between care team members, with frequently updated notes being also frequently viewed. Moreover, frequently updated notes commonly originated from one of the intensive care units. This indicates that the electronic resident handoff note is emerging as a new method for disseminating patient information, in addition to facilitating transitions of care. Given that there is a large proportion of mismatched utilization of handoff notes, more work can be done to help teams reach a consensus about using handoff notes as an alternative information source.

Acknowledgements
The research described was supported by the T15LM007079 grant from the National Library of Medicine.

References
7. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: an evaluation of information
doi:10.1016/j.ijmedinf.2014.06.011

doi:10.1542/peds.2006-0367


doi:10.1136/jamia.2010.008441


doi:10.1016/j.ijmedinf.2010.09.009

Causal Phenotype Discovery via Deep Networks

David C. Kale, MS,1,2 Zhengping Che,1 Mohammad Taha Bahadori, PhD,1 Wenzhe Li, MS,1 Yan Liu, PhD,1 Randall Wetzel, MD2
1 University of Southern California, Los Angeles, CA
2 Whittier Virtual PICU, Children's Hospital Los Angeles, Los Angeles, CA

Abstract

The rapid growth of digital health databases has attracted many researchers interested in using modern computational methods to discover and model patterns of health and illness in a research program known as computational phenotyping. Much of the work in this area has focused on traditional statistical learning paradigms, such as classification, prediction, clustering, pattern mining. In this paper, we propose a related but different paradigm called causal phenotype discovery, which aims to discover latent representations of illness that are causally predictive. We illustrate this idea with a two-stage framework that combines the latent representation learning power of deep neural networks with state-of-the-art tools from causal inference. We apply this framework to two large ICU time series data sets and show that it can learn features that are predictively useful, that capture complex physiologic patterns associated with critical illnesses, and that are potentially more clinically meaningful than manually designed features.

1 Introduction

The increasing volume, detail, and availability of stored digital health data offers an unprecedented opportunity to learn richer, data-driven descriptions of health and illness.1 This principle has driven the rapid development of computational phenotyping, a broad field encompassing a variety of efforts to apply modern computational methods to increasingly large and complex medical data sets. These efforts are further unified by a common goal: to not only build successful predictive models but also to learn and recognize clinically meaningful descriptors of health. Computational phenotyping has attracted many researchers in machine learning and data mining, who apply many different techniques (ranging from Gaussian processes to tensor factorization) to extract meaningful representations from a variety of data sources and types (e.g., clinical time series, text, event counts, etc.).1,2,3,4,5,6,7

Some of the most interesting phenotyping work involves phenotype discovery, i.e., learning latent representations that capture structure in data and may represent real patterns of illness. Such techniques have the potential to become powerful tools for clinical research. For example, Schulam, Wigley, and Saria used a Bayesian framework to find patient clusters that may represent previously unknown subtypes of Scleroderma.6 However, phenotype discovery from physiologic data poses significant challenges because human physiology is immensely complex and nonlinear, and illnesses often span a large number of body systems and processes. Acute respiratory distress syndrome (ARDS), for example, involves an acute failure of the respiratory system but may manifest with circulatory, excretory, and even neurological symptoms. What is more, its causes range from injury from inhalants to infection to overdose of antidepressants.8 This complexity suggests that such diseases have many latent factors of variation, and modeling and disentangling these factors is essential to successful analysis.9 One potential solution to this problem is deep learning (e.g., multilayer neural networks), which has led to major breakthroughs in speech recognition10 and computer vision11.

However, despite rapid advances in methods12 and software,13,14,15,16 there has been comparatively little formal research on interpretation of deep learning architectures.17,18 This becomes an especially critical question when applying deep learning to health and medicine. Neural networks are classically viewed as "black box" models that may achieve high predictive accuracy but are uninterpretable by humans and unsafe for use on clinical problems. One solution for unraveling the complex representations produced by deep learning is to apply ideas and tools from causal inference.19 Feed-forward architectures are in fact directed acyclic graphs (DAGs), in which inputs cause higher layer activations, which in turn cause outputs. Thus, they may be thought of as causal models, which makes them amenable to causal analysis. This likewise provides an motivation for phenotype discovery, in which we seek to learn representations that are potential causes of outcomes, rather than merely correlated or predictive.

There is a growing body of research on discovering causal relationships among variables from observational data under a variety of assumptions and settings.20,21,22 In particular, we can identify potential causal relationships among the variables if noise is not Gaussian distributed.21,22 This is frequently the case for neural networks with, e.g., sigmoid hidden units and binary outcomes. In this paper, we present a first step toward automatic discovery of causal phenotypes and for cracking open the black box of neural networks, making them more readily applicable to medical
data. Our frame is a two-stage process. First, we use a simple deep neural network architecture to learn latent representations of physiology from clinical time series. Then we use a state-of-the-art causal inference algorithm called Pairwise LiNGAM\(^22\) to analyze the relationships between these learned phenotypes and patient outcomes and diseases of interest. Finally, we use common deep learning heuristics to visualize and interpret the learned phenotypes. We show that this algorithm discovers intuitive patterns of physiology known to be associated with acute illnesses. We also propose an informal causality-based framework for measuring the causal power of learned representations.

### 1.1 Related Work

While a young field (at least by that name), computational phenotyping is advancing rapidly, spurred on by the increased adoption of electronic health records (EHRs) and the growing interest from data mining and machine learning researchers. There is already a large body of excellent research, much of it published in just the last five years or so.

One popular approach to computational phenotyping is to construct a large multi-dimensional array (i.e., matrix or tensor) view of clinical data and then apply dimensionality reduction or feature selection techniques to learn a lower-dimensional set of latent bases that can be treated as phenotypes. Each basis can be seen as a combination of observations (from the original tensor), while each patient becomes a sparse set of coefficients representing her projection onto the phenotype bases. Such work can be seen as a much more powerful generalization of classic techniques such as principal components analysis. Two primary examples of this paradigm include Zhou, et al.,\(^4\) and Ho, Ghosh, and Sun,\(^5\) which apply such frameworks to outpatient disease data and medicare claims, respectively, with very interesting results. Such an approach could be applied to physiologic time series, as well.

An alternative phenotyping paradigm includes probabilistic models, which assume a generative process and then fit the model to data using, e.g., maximum a posteriori inference. Such models can be robust to uncertainty, noise, and some types of missing values and are often interpretable. Marlin, et al., used a Gaussian mixture model with a temporal smoothness kernel prior to discover meaningful physiologic patterns (or physiomes) in multivariate time series from acute care settings similar to ours.\(^1\) Saria, et al., proposed a Time Series Topic Model (TSTM) that can learn bag-of-AR (autoregressive linear models) representations from dense time series (e.g., ECG waveforms) and has been used to develop a novel severity of illness score for neonatal patients.\(^7\) Schulam, Wigley, and Saria recently proposed the Probabilistic Subtyping Model, a Bayesian framework that combines splines and Gaussian processes to cluster longitudinal data from chronic Scleroderma patients and discover potentially novel subtypes.\(^6\)

To our knowledge, work by Lasko, Denny, and Levy represents one of the first applications of modern deep learning to clinical time series.\(^3\) They train stacked autoencoders on 30-day windows of uric acid readings to learn features that are competitive with expert-designed features for classifying gout versus leukemia. They handle irregular, biased sampling by warping their time series and then sampling from a fitted Gaussian process. This framework successfully learns time series features that are both visually intuitive and useful for discerning the two phenotypes. Kale, et al., and Che, et al., recently demonstrated that neural networks (unsupervised and supervised) can be used to discover and detect interpretable subsequences in multivariate physiologic time series that are useful for classifying acute illnesses.\(^23,24\) Given that deep learning approaches have achieved breakthrough results in language modeling,\(^25\) speech recognition,\(^10\) and music transcription,\(^26\) we expect similar results (with time and effort) in health and medicine.

In deep learning research, feature analysis is often secondary to, e.g., prediction performance, and focuses on visualization. Strategies include sampling from generative models and optimizing (using, e.g., stochastic gradient ascent) over inputs rather than parameters.\(^17\) Each method has strengths and weaknesses (e.g., simplicity, computational efficiency, local optima), but they share several properties: they work best for data that are easily interpreted by human beings (e.g., images\(^27\)); they employ heuristics and approximations; and they analyze each hidden unit independently. However, recent research has begun to provide a more rigorous understanding of the representations learned by deep architectures. Szegedy, et al., showed that the semantics encoded by hidden unit activations in one layer are preserved when projected onto random bases, instead of the next layer's bases.\(^28\) This implies that the practice of interpreting individual units can be misleading and that the behavior of deep models may be more complex than previously believed.\(^29\)

One solution for unraveling the complex representations produced by deep learning is to apply ideas and tools from causal inference.\(^19\) Chalupka, Perona, and Eberhart recently proposed a theoretical framework that reformulates image classification as a causality problem (i.e., an image causes an agent to label it as a 7 or 9) and uses active learning to perform interventions that can separate causal features from spurious correlations.\(^30\) This idea offers a partial solution to the problems described above\(^28\) but requires the ability to perform interventions that may not be possible in clinical data analysis. Alternatively, there is a growing body of research on discovering causal relationships among variables...
from observational data under a variety of assumptions and settings,\textsuperscript{20,21,22} which we discuss in detail later.

2 Methods

In this section we describe our two-stage framework for discovery and analysis of causal phenotypes from clinical time series data uses deep neural networks. We first describe the background of feature (i.e., phenotype) extraction from time series in Section 2.1. We then demonstrate how deep neural networks can be used to perform both unsupervised (Section 2.2) and supervised (Section 2.3) discovery of latent representations of physiology (i.e., phenotypes) from clinical time series. Finally, we show in Section 2.4 how state-of-the-art causal inference algorithms can be used to analyze the learned phenotypes and to identify potential causal relationships between phenotypes and critical illness.

2.1 Background: feature extraction from time series

Given a multivariate time series with $P$ variables and length $T$, we can represent it as a matrix $X \in \mathbb{R}^{P \times T}$. We denote the time series of the $p$th variable as a row vector $x_{p, \cdot} \in \mathbb{R}^T$ and the $t$th time as a column vector $x_{\cdot, t} \in \mathbb{R}^P$.

A feature map for time series $X$ is a function $f : \mathbb{R}^{P \times T} \rightarrow \mathbb{R}^D$ that maps $X$ into a $D$-dimensional feature space, which can be used for machine learning tasks like classification, segmentation, and indexing.\textsuperscript{31} In a medical context, we can think of features as phenotypes. These can take the form of extreme measurements (as in severity of illness scores), thresholds, or important patterns. In multivariate time series, features become increasingly complex to design, so automated feature discovery is an attractive proposition. Given the recent success of deep learning in a variety of applications, it is natural to investigate its effectiveness for feature learning from clinical time series data.

![Image](image1.png)

Figure 1: Deep neural networks for phenotyping from clinical time series.

2.2 Deep unsupervised autoencoders for phenotyping clinical time series

We explore several deep neural network architectures for automatic discovery and detection of important physiologic patterns. We begin with a simple denoising autoencoder (DAE).\textsuperscript{32} This is a one layer unsupervised model that simultaneously learns paired encoding and decoding functions, similar to sparse coding but easier to optimize and incorporate into deep architectures. Figure 1a shows a simple illustration of the DAE.

We encode and decode $x$ using rules similar to making a prediction using a logistic function:

$$h = g(Wx + b), \quad \hat{x} = g'(W' h + b')$$

where $h \in [0, 1]^D$ is the latent representation, $\hat{x}$ is the reconstruction, and $g$ and $g'$ are elementwise nonlinearities (a common choice is the sigmoid function $g(z) = 1/(1 + \exp\{-z\}$). The choice of $g'$ depends on the type of input $x$. As described later, in this work we scale all variables to fall between 0 and 1, so we can use a sigmoid for decoding as
well. As is typical, we also tie the weights, letting \( W' = W^T \). Finally, in DAEs we actually add random corruption to the input before encoding, sampling \( \hat{x} \sim P_{corr}(\hat{x}|x) \). In our case, \( P_{corr} \) applies a binary masking to \( x \), zeroing out each entry independently with some probability \( p \).

We train the weights by minimizing the reconstruction loss for each training example. For \([0, 1]\) inputs, we use cross entropy loss

\[
\mathcal{L} = - \sum_{d=1}^{D} (x_d \log \hat{x}_d + (1 - x_d) \log(1 - \hat{x}_d)),
\]

where \( x_d \) is the \( d \)th dimension of \( x \). Note that for a DAE, \( \hat{x} = g'(W^T g(W \hat{x} + b) + b') \), i.e., the reconstruction of the corrupted input \( \hat{x} \). We use standard (stochastic) gradient methods to minimize the reconstruction error with respect to \( W, b \) and \( b' \) for all training examples.

We can construct a deep autoencoder by stacking multiple DAEs, forming a stacked denoising autoencoder (SDAE).\(^{33}\) SDAEs are typically trained using greedy layer-wise training,\(^{34}\) as shown in Figure 1b. Once the weights for layer \( \ell \) have been trained, we can map each training example into its feature space, producing \( h^{(\ell)} \). This then becomes the input for training the \((\ell+1)\)th layer. Note that \( h^{(0)} = x \), the input.

In our setting, \( x = \text{vec}(X) \), a vectorization of our \( P \times T \) time series. We can then use an SDAE of any number of layers as a feature map \( f \) for time series, as shown in Figure 1c. Each element of \( h^{(\ell)} \) is a nonlinear function of the SDAE’s inputs, meaning that it can capture complex correlations across both time and variables. This makes it well-suited tool for phenotyping from clinical time series, especially when working with relatively small data sets with few or unreliable labels.\(^3\)

### 2.3 Deep supervised neural networks for phenotyping clinical time series

We can convert a deep autoencoder into a deep feed-forward neural network by adding an additional output layer to make predictions, as shown in Figure 1c. For binary classification, we typically use a sigmoid nonlinearity applied to a linear activation, i.e., a logistic regression:

\[ y_k = \sigma(\beta^T_k h^{(\ell)}) \]

where \( y_k \) is the \( k \)th output unit and \( h \) are the hidden unit activations (we omit the bias for brevity). Neural networks lend themselves naturally to multi-output prediction problems (also called multi-label or multi-task learning), and training such neural networks can often improve prediction performance by enabling the neural network discover shared features that are useful across a range of tasks. In a medical context, this approach can be used to train a single model to predict multiple outcomes or diagnoses.

A neural network with \( L \) hidden layers and an output layer has hidden layer parameters \( \{ (W^{(\ell)}, b^{(\ell)}) \}_{\ell=1} \) and output parameters \( \{ \beta_k \}_{k=1}^K \) for \( D^{(L)} \) hidden units in the \( L \)th layer and \( K \) outputs. For \( K \) binary classification tasks, the loss function during supervised training also uses cross-entropy but with the true labels (vs. reconstruction, as in the SDAE):

\[
\mathcal{L} = - \sum_{k=1}^{K} \left( y_k \log \sigma(\beta^T_k h^{(L)}) + (1 - y_k) \log(1 - \sigma(\beta^T_k h^{(L)})) \right)
\]

where \( h^{(L)} = g(W^{(L)} h^{(L-1)} + b^{(L)}) \) and \( h^{(0)} = x \). Again, we minimize the loss with respect to all model parameters using (stochastic) gradient descent and backpropagation.

### 2.4 Discovery of causal phenotypes from clinical time series

One of the main advantages of deep learning is its ability to disentangle factors of variation that are present in the data but unobserved.\(^9\) This makes subsequent learning (i.e., training a classifier) much easier since it counteracts the curse of dimensionality.\(^{35}\) In addition, knowledge about one factor usually improves estimation about another.\(^{36}\)

However, it can often be difficult to analyze and understand the learned latent representations and to understand whether the model is learning truly important relationships or spurious correlations. One way to explore and demonstrate this fact is to perform causal analysis of the features extracted by the hidden layers of a deep neural network. Disentangled representations should have clearer and stronger causal relationships with other variables of interest (e.g., mortality) than raw outputs and other choices of features. Additionally, causality is of primary interest in medicine and health, especially if analytics will contribute to decisions about treatment and care, which can significantly impact patient lives and outcomes. Thus, discerning correlation from true causal relationships is of vital importance.
Given a set of features denoted by $\mathbf{h} \in \mathbb{R}^{D(L)}$ and a response variable $y$, we investigate the causal relationship between each feature $h_j$, $j = 1, \ldots, D(L)$, and the response variable $y$. There are two options: either the direction of the edge is from feature to the response variable $h_j \rightarrow y$ or vice versa $h_j \leftarrow y$. We are interested only in the former case where the features are causally predictive of the response variable. Thus, we need to use a causality discovery procedure to find the direction of causation between the features and the response variable.

Classic causal inference algorithms require a set of causal priors to be available for the variable to be able to cancel out the impact of spurious causation paths. While we often do not have such priors available for our outputs, we can still identify causation among the variables if they are not distributed according to a Gaussian distribution. Binary labels (e.g., mortality prediction) satisfy the requirements of many causal inference frameworks. We apply a state-of-the-art causal inference algorithm Pairwise LiNGAM, based on DirectLiNGAM, in order to discover the causal edges between each feature and response variable. The key idea of this algorithm is to compute the likelihood ratio of the two models $h_j \rightarrow y$ and $h_j \leftarrow y$ for $j = 1, \ldots, D(L)$ and select the direction that makes the log-likelihood ratio positive. In particular, we have

$$R = \frac{1}{n} \mathcal{L}(h_j \rightarrow y) - \frac{1}{n} \mathcal{L}(h_j \leftarrow y) \quad \Rightarrow \quad \begin{cases} h_j \rightarrow y & \text{if } R > 0, \\ h_j \leftarrow y & \text{if } R < 0. \end{cases}$$

where $n$ denotes the number of observations. The log-likelihood values are computed using the non-parametric entropy estimation techniques. Pairwise LiNGAM requires that the two variables be non-Gaussian distributed, which makes it especially useful for analyzing deep neural networks with nonlinear activation functions in hidden layers and logistic outputs. This is the case in our setting.

It is important to emphasize that causal inference algorithms do not necessarily select those features that are most correlated with the response or most useful in predicting it. Our goal in causal analysis, rather, is to discover a subset of features that are the best candidates to be true causes of the response and which may provide insight (not necessarily more predictive power).

Next we propose an informal method for quantifying the causal power of features (derived or learned). After learning the causal features, we follow the recommendation, we fit a logistic regression model to the features selected by the causality discovery algorithm as follows:

$$\hat{\alpha}, \hat{\alpha}_0 = \arg \max_{\alpha, \alpha_0} \left\{ \sum_{i=1}^{n} [y_i \log \sigma(\alpha^\top \hat{\mathbf{h}}_i + \alpha_0) + (1 - y_i) \log(1 - \sigma(\alpha^\top \hat{\mathbf{h}}_i + \alpha_0))] \right\}$$

where $\hat{\mathbf{h}}$ represent the set of features selected by the causality discovery algorithm and $\alpha, \alpha_0$ denote the prediction vector and the intercept, respectively. We treat the resulting weights as the magnitude of each variable's causal relationship. Finally, we use the $L_2$ norm of the regression coefficient vector $\|\alpha\|_2$ to quantify the overall causal power of the features being analyzed. We can use this to compare the causal power of different representations.

3 Experiments

In order to demonstrate the effectiveness of our framework, we performed a series of phenotype experiments using two clinical time series data sets collected during the delivery of care in intensive care units (ICUs) at large hospitals. After describing our data and experimental set up, we briefly present quantitative results for several classification tasks, in order to demonstrate the predictive power of features discovered by neural networks (Section 3.1). Then in Section 3.2, we apply causal inference tools to the learned features in order to discover the most clinically meaningful features and to analyze the quality of the learned phenotypes. We also provide example visualizations of causal features learned by neural networks that capture clinically significant physiologic patterns.

**Physionet Challenge 2012 Data.** The first data set comes from PhysioNet Challenge 2012 website which is a publicly available collection of 8000 multivariate clinical time series from one ICU and three specialty units, including coronary care and cardiac and general surgery recovery units. As with the competition, we focus on mortality prediction (from the first 48 hours of each episode) as our main prediction task. This is a challenging problem: no competition entry scored precision or recall higher than 0.53. We used both Training Subsets (A and B) for any unsupervised training but only the labeled Training Subset A for supervised training and evaluation (to our knowledge, labels are not available for Subset B). Each episode is a multivariate time series of roughly 48 hours and containing over 30 variables. While each episode also has a variety of static variables available (e.g., age, weight, gender), we focus our experiments
on just the time series. In all supervised learning experiments, we use label stratified 10-folds cross validation when estimating performance scores.

**PICU Data.** The second data set consists of ICU clinical time series extracted from the electronic health records (EHRs) system from Children's Hospital LA (CHLA), previously described in Marlin, et al, and Kale and Che, et al. The original data set includes roughly ten thousand episodes of varying lengths, but we exclude episodes shorter than 24 hours, yielding a data set of roughly 8500 multivariate time series of thirteen physiologic variables. Each episode has zero or more associated diagnostic codes from the Ninth Revision of the *International Classification of Diseases*, (ICD-9). We aggregate the five-digit ICD-9 codes according to the standard seventeen broad category codes (e.g., 460-519 for respiratory diseases) and supplementary V and E groups. We then treat predicting each category code as a distinct binary classification task. The sparse multi-label nature of these data prevents us from applying cross-validation; we instead create five 80/20 random splits of the data, ensuring that each split has a minimum number of positives examples for each ICD-9 label.

**Preprocessing.** We perform three steps of preprocessing to both data sets before analysis. First, we scale each variable to a $[0, 1]$ range. Where variables have known ranges (e.g., Total Glasgow Coma Scale or binary variables), we use these. Otherwise, we treat the 1st and 99th percentiles of all measurements of a variable as its minimum and maximum values. Outliers are truncated to 0 or 1. This is applied to both time series and static variables. Next, we resample all time series to a fixed hourly sampling rate using a simple bucketing procedure: we divide each time series into 48 non-overlapping hour-long windows. When a window includes more than one measurement, we take the mean. Where this creates missing values, we propagate forward the previous measurement. This makes a reasonable assumption that each time series is relatively stable and that important changes are observed and recorded. Finally, we handle entirely missing time series (e.g., a patient may have zero measurements of end-tidal CO$_2$ if she is not ventilated) by imputing a `normal` value. For variables without known normals, we use the median of all measurements in the data set. This strategy the fact that missing time series are typically not missing-at-random but rather are missing because clinical staff decided not to measure a particular variable. Often this is because they also assume it is normal.

**Neural network training.** We implemented all neural networks in Theano as variations of a multilayer perceptron with 3-5 hidden layers (of the same size) of sigmoid units. The input layer has $PT$ input units for $P$ variables and $T$ time steps, while the output layer has one sigmoid output unit per label. We initialize each neural network by training it as an unsupervised stacked denoising autoencoder (SDAE). We found this helps significantly because our data sets are relatively small and our labels are quite sparse. We use minibatch stochastic gradient descent to minimize cross-entropy loss during unsupervised pretraining and logistic loss during supervised finetuning. We use ten-fold cross validation, and both neural networks and classifiers are not trained on the test folds. Additionally, we use grid search and one training fold to tune parameters (e.g., the strength of the L1 penalty).

### 3.1 Classification performance

We first present a quantitative evaluation of the predictive performance of different types of features, both hand-designed and learned, on the Physionet Challenge 2012 data set. To ensure a fair comparison, we use the same type of classifier in all experiments: a linear support vector machine (SVM) with hinge loss and a designed and learned, on the Physionet Challenge 2012 data set. To ensure a fair comparison, we use the same type of cross validation, and both neural networks and classifiers are trained on the test folds. Additionally, we use grid search and one training fold to tune parameters (e.g., the strength of the L1 penalty).

We do not use our neural networks to make predictions. We select the strength of the L1 penalty by performing a grid search over the range $[10^{-2}, 10^{2}]$ and choosing the value that maximizes the Area Under the Precision-Recall Curve (AUPRC) on a held-out subset of our training data. We report both Area Under the Receiver Operator Curve (AUROC) and AUPRC, as well as the Precision when Recall is 90%. All three metrics are more robust to the class imbalance of our label than accuracy and give us an idea of the trade-off between false negatives and false positives.

Our baselines include the raw data and hand-designed features that capture the extremes, central tendency, variation, and trends within the entire time series. While relatively simple from a machine learning perspective, these features are often quite effective for clinical predictive modeling and similar to those used in classic severity of illness scores. Table 1 shows the mortality prediction performance for the Physionet Challenge 2012 data for our best-performing baselines and neural network features. We see that features learned using a 3-layer neural network beat the raw data fairly substantially and are competitive with the hand-designed features. Given the success of neural networks in other domains, it is somewhat disappointing that the learned features do not beat the hand-engineered features soundly. However, we offer several observations to temper this disappointment: first, we invested minimal time in tuning hyperparameters of the neural network, including hidden layer sizes, learning rate, and early stopping criteria. Additional experiments (not reported here) suggest that our neural network underfit the training data due to insufficient size and training epochs.
Second, as shown in the bottom two rows of Table 1, we found that we could substantially improve performance by combining the neural net and hand-engineered features. This suggests that in fact the neural net learns features that contain information that does not overlap that captured by the hand-engineered features and may capture different physiologic patterns. What is more, as shown in Section 3.2, they have increased causal power.

Interestingly, we found minimal difference between the performance of features learned using unsupervised and supervised neural networks. We speculate that this has two causes. The principle reason, we speculate, is the class imbalance in our labels, which we did not attempt to handle in any way during neural network training. Second, mortality prediction from early admission data is a difficult problem, and it may not be possible to do substantially better (our results are similar to those from the competition).

### 3.2 Causal analysis

Next, we perform causal inference on the hand-designed and learned features using the framework described in Section 2.4. Table 2 shows the per-feature causal power (i.e., the $L_1$ norm of the coefficient vector divided by the number of features with nonzero weights) of the raw data, hand-designed features, and neural network features for Acute Respiratory Distress Syndrome (ARDS) in the PICU data. We see the the neural network features have, on average, a much larger magnitude than either of the baselines.

<table>
<thead>
<tr>
<th></th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Precision@90%Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Time Series (R)</td>
<td>0.786848 ± 0.028957</td>
<td>0.407419 ± 0.042878</td>
<td>0.221303 ± 0.017106</td>
</tr>
<tr>
<td>Hand-designed Features (H)</td>
<td>0.828652 ± 0.021065</td>
<td>0.467742 ± 0.047852</td>
<td>0.259324 ± 0.049400</td>
</tr>
<tr>
<td>NNet(R,3)</td>
<td>0.820760 ± 0.021021</td>
<td>0.444315 ± 0.032367</td>
<td>0.255792 ± 0.030306</td>
</tr>
<tr>
<td>H+R</td>
<td>0.822907 ± 0.018251</td>
<td>0.438160 ± 0.035444</td>
<td>0.255608 ± 0.031871</td>
</tr>
<tr>
<td>H+NNet(R,3)</td>
<td>0.845015 ± 0.016525</td>
<td>0.486791 ± 0.047373</td>
<td>0.291411 ± 0.033500</td>
</tr>
</tbody>
</table>

Table 1: Classification performance on the Physionet Challenge 2012 data set. We report mean and standard deviation (across 10 folds) for each metric. We use the following abbreviations: $R$: raw time series, $H$: hand-designed features, $NNet(I,L)$: L-layer neural network with input $I$

Figure 2 shows visualizations of two the significant physiologic patterns learned by the neural networks. For each we used causal inference to discover the subset of features with the strongest causal relationship with our outcome of interest. Then we found the 50 input subsequences with the highest activations in those units and plotted the mean trajectories for some or all physiologic variables. Figure 2a visualizes features that were found to be causally related to the ICD-9 circulatory disease category from the PICU data. We see these features detect highly elevated blood pressure and heart rate, as well as depressed pH. The features also detect elevated end-tidal CO2 (ETCO2) and fraction-inspired oxygen (FIO2), which likely indicate ventilation and severe critical illness. Interestingly, these features also detect elevated urine output, and thus it is not surprising that these features are also correlated with diagnostic labels related to urinary disorders. Figure 2b visualizes the First-48-hour physiologic patterns detected by features that are causal of mortality in the Physionet Challenge 2012 data.

### 4 Discussion and Conclusion

We have presented a simple, two-stage framework for discovering latent phenotypes from clinical time series that have strong causal relationships with patient outcomes and critical illness. Our framework combines feature learning using neural networks with causal inference tools to discover latent phenotypes that are causally predictive of clinical outcomes. While our results are preliminary, we believe that this general line of research will help us discover more clinically meaningful representations of health and illness and to eventually develop tools for automatic discovery of causal phenotypes.
(a) 12-hour causal phenotype for ICD-9 circulatory disease category (390-459), learned from the PICU data.

(b) 48-hour causal phenotype for mortality, learned from the Physionet Challenge data.

Figure 2: Causal features learned from ICU time series.

5 Acknowledgments

David Kale was supported by the Alfred E. Mann Innovation in Engineering Doctoral Fellowship, and the VPICU was supported by grants from the Laura P. and Leland K. Whitter Foundation. Mohammad Taha Bahadori was supported by NSF award number IIS-1254206. Yan Liu was supported by NSF IIS-1134990 and IIS-1254206 awards. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agency, or the U.S. Government.

References


Initial Readability Assessment of Clinical Trial Eligibility Criteria

Tian Kang, BS, Noémie Elhadad, PhD, Chunhua Weng, PhD
Department of Biomedical Informatics, Columbia University, New York, NY 10032

Abstract

Various search engines are available to clinical trial seekers. However, it remains unknown how comprehensible clinical trial eligibility criteria used for recruitment are to a lay audience. This study initially investigated this problem. Readability of eligibility criteria was assessed according to (i) shallow and lexical characteristics through the use of an established, generic readability metric; (ii) syntactic characteristics through natural language processing techniques; and (iii) health terminological characteristics through an automated comparison to technical and lay health texts. We further stratified clinical trials according to various study characteristics (e.g., source country or study type) to understand potential factors influencing readability. Mainly caused by frequent use of technical jargons, a college reading level was found to be necessary to understand eligibility criteria text, a level much higher than the average literacy level of the general American population. The use of technical jargons should be minimized to simplify eligibility criteria text.

Introduction

As the gold standard for generating the most rigorous medical evidence regarding the effectiveness and efficacy of new medical therapies, clinical trials are fundamental for advancing medical research and public health. However, recruitment has remained the biggest persistent obstacle. With the pervasive Internet access, the rise of web-based patient-trial matching and patient-screening methods presents great opportunities to overcome recruitment barriers. There is an increasing need to engage the general public in participating in clinical trials through “active matching,” where patients or health consumers are empowered to review clinical trial eligibility criteria and select clinical trials to participate in. According to the search logs of popular online clinical trial search engines, in addition to researchers, patients and health consumers have been increasingly searching and browsing clinical trials.

The success of active patient-trial matching hinges on the readability of clinical trial summaries, particularly the eligibility criteria language that clinical trial seekers use for determining their match to a trial. Because of the different professional training levels between patients and the researchers who wrote the clinical trial summaries, we hypothesize that required reading level of the trial texts are higher than the health literacy of out-of-domain readers. In particular, eligibility criteria, which determine the potential match for a patient, contain complex, specific content primarily written for researchers. In this paper, we examine the readability of clinical trial eligibility criteria.

Text readability assessment is an established field of research. As defined by Dale and Chall et al., readability is “the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at optimal speed, and find it interesting”. Most readability measures have been established as assessment of shallow features of text, such as the number of words per sentence and the number of syllables per word. The SMOG readability assessment formula, which measures the percentage of polysyllables in texts, is a popular method for semantic assessment. These widely adopted assessments are easy to compute on any text but lack the ability to represent the complexity of a text accurately, as this requires the use of both semantic and syntactic features. Measures of readability, which do assess both semantic features and syntactic features, have also been adopted to achieve more accurate assessment results. The New Dale-CHALL formula uses a combination of average sentence length and the proportion of the words in the text deemed “difficult” based on a 3000-word list containing what have been identified as the most commonly used English words. As an added benefit, Natural Language Processing (NLP) enables automated readability assessment. For instance, Yngve measures the depth of nodes in the parse tree using NLP methods as the indicator of structure complexity. These well-accepted methods for readability assessment have been validated in general reading materials and have been in use for years. Feng et al. performed a readability measurement for people with intellectual disabilities (ID) using a linear regression model to fit a set of cognitively motivated features. Methods like Support Vector Machine were also popular in this kind of research. However, supervised learning approaches constantly suffer from one big limitation, the need for large-scale labeled corpora.

Compared to what is known about general English’s readability, our understanding of biomedical text’s readability, especially that of clinical trial summaries, remains limited. Many studies of medical texts focus on online health-related materials and evaluate if these texts’ readability meets the recommendations of NIH, AMA and USDHHS.
standards (less than 6 grade) using general readability formulas. A review of all readability and comprehension instruments used for print and web-based cancer materials concluded that although readability formulas are predictive measures of text comprehensibility, they have been criticized for “only relying on word and sentence factors and for ignoring possible effects for reader motivation, design, and graphics on readability and comprehension”. Kim et al. developed a new health text-specific readability measurement, which is based on the distance of text features from the health related material to the predefined “easy” and “difficult” documents. It takes both syntactic features and semantic features into consideration and generates a sum of those distances as the indicator of readability. This method has been applied to several studies since its inception.

Since clinical trials constitute a critical part of the biomedical domain, the readability of clinical trials has started to receive attention. For example, Wu et al. conducted a readability assessment for description text in clinical trials from ClinicalTrials.gov using both general purpose and medical specific readability assessment measures. They concluded that the descriptions were the most difficult to read when compared to other corpora e.g. electronic health records, MedlinePlus Health Topics articles. Ross et al. conducted the first formal analysis of computer-interpretations of eligibility criteria complexity in clinical trials in order to facilitate phenotype studies. They randomly selected 1000 clinical trials from ClinicalTrials.gov and manually analyzed their complexity and semantic patterns. The results concluded that 93% of these free-text criteria were mainly comprehensible for professionals and 85% had significant semantic complexity for computational representations. Still, little is known about the readability of eligibility criteria and its impact on clinical trial recruitment and evidence adoption. Aiming to fill this knowledge gap and guide future improvement of consumer-facing clinical trial search engines, this paper contributes the first comprehensive readability assessment of clinical trial eligibility criteria text. In this study, we measured the human reading level necessary to understand the eligibility criteria in ClinicalTrials.gov. We measured both general readability and health-domain specific readability of eligibility criteria and discussed the implications of our findings for clinical trial designers and clinical trial search engine developers.

**Methods**

**Datasets**

Three corpora were used for comparison in this project. We retrieved clinical trial summaries from the world’s largest clinical trial registry, ClinicalTrial.gov, as our target corpus. To generate a better understanding of the results, we also performed assessments on two health-related corpora, i.e., Reuters News and PubMed. Reuters News is intended for lay people. PubMed is intended for technical readers. Samples taken from the three corpora are shown in Table 1. We select a set of disease topics to include disease-specific descriptions from the three corpora. We downloaded the flat file of the database from the ClinicalTrials.gov website using their API and extracted trial summary text for the same list of health topics in two health-related corpora (e.g., Alzheimer, Type 2 Diabetes Mellitus, and Cardiovascular Diseases). We matched these disease topics to the condition field of clinical trial summaries supplied by ClinicalTrials.gov. We extracted 120,977 clinical trials on the aforementioned health topics. Then we collected 3,144 Reuters stories along with their corresponding PubMed articles.

**Table 1. Sample text on the topic of “lung cancer” from the three corpora**

| PubMed articles | ...Marijuana smoke contains many of the same constituents as tobacco smoke, but whether it has similar adverse effects on pulmonary function is unclear...The Coronary Artery Risk Development in Young Adults (CARDIA) study, a longitudinal study collecting repeated measurements of pulmonary function and smoking over 20 years...
 |
| --- | --- |
| Reuter News | ... A few hits on the bong now and then don't seem to have any detrimental effects on lung health, suggests a new study. Researchers found that multiple measures of lung function actually improved slightly as young people reported using more marijuana ...
 |
| Eligibility Criteria | Ages Eligible for Study: 18 Years and older  
Genders Eligible for Study: Both  
Inclusion Criteria:  
• Pathologically confirmed, by biopsy or cytology, non-small cell lung carcinoma diagnosed within 3 months prior to study enrollment.  
• T1, N0, M0 or T2, N0, M0.  
• ...  
Exclusion Criteria:  
• Evidence of distant metastasis (M1) and/or nodal involvement (N1, N2, N3). |
The new Dale-Chall formula was also used. This measurement is considered more accurate than the original formula as it is based on the use of “easy” and “difficult” words rather than syllables or simply letter counts. It labels words found on an established word list containing 3000 statistically common-used English words as “easy” word if the word is not included in the word list, then this word is labeled as “difficult” and a corresponding penalty is added to the final score. The final score can then be mapped to the grade level and the required grade reading level for readers can be determined. The mapping from SMOG grade to U.S. standard educational level and the mapping from final score to education levels are shown Table 2.

The calculation method is as follows:

Raw score = \(0.1579 \times \) percentage of “difficult words” + average sentence length

If percentage of “difficult words” >5%

Adjusted score = Raw Score + 3.6365,

Otherwise Adjusted score = Raw score

Table 2. Mapping of Readability Levels across Standards
2. Term familiarity prediction

We hypothesized that the technical terminology used in eligibility criteria might be one of the major obstacles preventing lay readers from fully understanding the content. Thus, a health-domain specific method published by Elhadad [2] was applied to evaluate the familiarity of words used in EC texts based on word frequency. The words with high frequencies in “easy” texts are usually found to elicit a higher recognition than words with lower frequencies. We know that the Reuter Health news stories are targeted at lay readers, thus when a word shows a high frequency (in its all morphological variants) in this corpus, we can define it as “familiar” (“easy”) for lay readers. In contrast, the words with high frequencies in PubMed articles, but rarely seen in the news stories, are more likely to elicit a higher requirement for professional knowledge, then this word tends to be more “unfamiliar” (“difficult”) for lay readers. Therefore, if the term usage in EC texts is more similar to Reuter News, then we could conclude that the terms used in EC texts are generally not difficult for lay readers; otherwise, the demands of the EC texts are more like technical papers. Figure 2 illustrates the 3-step approach we applied to evaluate term familiarity.

![Figure 2. Our analysis schema for familiarity evaluation of terms](image)

concept to contain several lexical variants mapping to the same CUI (e.g. CUI: C0027051, lexical variants: heart attack, myocardial infarctions). We generated a frequency distribution of all lexical variants for each CUI in each corpus and then applied Kullback–Leibler divergence to evaluate pairwise differences between Reuter news/ EC texts, and PubMed articles/ EC texts. When the difference between eligibility criteria texts and news was smaller, we defined the word usage in EC texts as relatively “easy” for lay readers. For example, for CUI C0027051, we determined the frequency of use of “heart attack” and “myocardial infarctions.” Commonly, the latter is a more technical term used in the health domain and so we expect a lower frequency in the Reuter news corpus, e.g. 0.01. Say we found that the frequency distribution of C0027051 in the news corpus was (0.99, 0.01); then we would calculate the frequency of the same CUI in the PubMed article. We expect that the frequency of “myocardial infarctions” would be much higher in technical papers than in the news corpus, e.g. (0.5, 0.5). When we apply KL divergence between the two distributions of the compared (news and article) and the test (clinical trials) corpus, those having similar distributions of word frequency can be detected. After manually filtering the concepts detected by HealthTermFinder that were not discriminative enough to be distinguished as technical or general text, such as C0014522 - "at any time," and also those not existing in our selected eligibility criteria texts, 90 UMLS concepts were finally selected. The KL divergence of Q from P is defined as:

$$D_{KL}(P||Q) = \sum P(i) \ln \frac{P(i)}{Q(i)}$$

3. Stratified analysis using clinical trial metadata

Considering that context may help with comprehension, we also took the complete text of each clinical trial into consideration. Besides the eligibility criteria, ClinicalTrial.gov has also archived the background information, the purpose, and some other detailed metadata of each clinical trial. To examine clinical trial metadata associated with readability of eligibility criteria, we classified the selected trials by their recruitment conditions and study results and chose SMOG and Dale–Chall scores as the indicators of required reading level. Further, we investigated clinical trial metadata that might be associated with a variance in reading levels. We tried to classify the trials using some other metadata (e.g., country that submitted the trial, the year that the trial was started). We selected two lexical features as indicators for the stratified analysis in that, 1) they are easy to compute, and 2) though the general readability formulas suffer from a lack of ability to precisely measure biomedical text readability, they are still able to indicate the discrepancies in required readability among different texts and enable stratified analyses. The text processing was carried out in Perl (v5.18.2), and the statistical analysis was performed using R 3.1.1.
Results

1. Features Evaluation

The overall readability measurement results are shown in Table 3. The first two columns reflect the results for the two standard corpora used for comparison with the EC texts. Examining the shallow features, we can see that generally PubMed articles are more complex than their corresponding news stories, using longer sentences and more complex words. This is consistent with our intuition about scientific papers. A more in-depth look at the syntactic features reveals that the PubMed articles tend to use more nouns in each sentence while Reuter news stories tend to use more verbs and subordinate clauses. This result also is consistent with the characteristics of each type of literature. The news style, used for news reporting in mass media, tends to use subject-verb-object construction and vivid, active prose, to inform the public compared to the corresponding scientific papers, which follow the style of technical writing, having longer sentences including a series of nouns or noun phrases. Also, at the word level, as a rule, journalists try not to rely on jargon and will not use a long word when a short one will work the same, because their readers are the masses and so have completed various levels of education. Scientists, however, are more likely to use professional terminology. These contrasting features are reflected in the lexical feature results – both methods used graded PubMed articles much higher than the news. For Reuter news, according to Table 1 and 2, both SMOG and Dale-Chall grades indicate a reading level requirement of around grade 13-14, reflecting the consistency of the two kinds of measurement. However, discrepancy regarding the reading grade level exists between the measurement results (grade 13-14, college) and the claim made by Reuter news (grade 12, high school). This can be explained by the findings in a review of readability16 -- SMOG usually results in a score one or two grade levels higher because it is based on stricter criteria for readability, like other similar assessment methods. Overall, PubMed articles tend to use longer and more unfamiliar terms than the news, requiring readers to have at least a college degree to fully comprehend them.

<table>
<thead>
<tr>
<th>Features</th>
<th>Reuter</th>
<th>PubMed</th>
<th>CT texts</th>
<th>EC texts</th>
<th>EC texts (95 CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shallow</td>
<td>aWPS</td>
<td>21.24</td>
<td>23.30</td>
<td>16.34</td>
<td>8.86 (8.83, 8.90)</td>
</tr>
<tr>
<td></td>
<td>aSPW</td>
<td>1.46</td>
<td>1.65</td>
<td>1.80</td>
<td>1.87 (1.87, 1.88)</td>
</tr>
<tr>
<td></td>
<td>aPPS</td>
<td>3.05</td>
<td>5.28</td>
<td>3.98</td>
<td>2.46 (2.45, 2.46)</td>
</tr>
<tr>
<td>Syntactic</td>
<td>aNP</td>
<td>11.36</td>
<td>12.97</td>
<td>7.12</td>
<td>4.09 (4.07, 4.11)</td>
</tr>
<tr>
<td></td>
<td>aVP</td>
<td>5.89</td>
<td>3.45</td>
<td>2.51</td>
<td>1.61 (1.60, 1.61)</td>
</tr>
<tr>
<td></td>
<td>aSBr</td>
<td>1.10</td>
<td>0.41</td>
<td>0.29</td>
<td>0.15 (0.15, 0.15)</td>
</tr>
<tr>
<td>Features</td>
<td>Dale-Chall</td>
<td>9.82</td>
<td>11.72</td>
<td>11.73</td>
<td>11.64 (12.08, 12.10)</td>
</tr>
</tbody>
</table>

* The 95% Confidence Interval for averages of the scores of eligibility criteria from the entire database were calculated using 1000 bootstrap

After ensuring that we had a full understanding of the results for the two standard corpora, we interpreted the corresponding results for the EC texts. Two kinds of EC texts were assessed here. As mentioned above, we were also concerned whether other parts of the trials documents might help readers to understand the EC of clinical trials besides the EC texts we were focusing on. Thus, the third column in Table 3 represents the assessment results for the complete document text of the clinical trials (including purpose, detailed description, eligibility criteria, etc.), which we refer to as “CT texts” here. Finally, the fourth column is the assessment results for the eligibility criteria text (“EC text”) alone. Compared to the Reuter and PubMed corpora, EC texts tend to use much shorter sentences (aWPS) and all 3 syntactic features indicate a particularly simple syntactic structure, especially in the EC text.

The reason for these results can be explained by the sample texts in Table 1. Eligibility criteria basically consist of short descriptions or the constraints for the characteristics of the target cohorts, and most of time they are not even a complete sentence, leading to particularly low results for aNP, aVP, and aSBr. In contrast, the syntax of the CT text is closer to that of the standard corpora and the assessment results with respect to lexical aspects (aSPW, aPPS, SMOG and Dale-Chall) further indicate that the CT texts are more similar to PubMed articles, indicating a college reading level requirement. However, discrepancy appeared between the results for the EC text in SMOG and those in Dale-Chall. The SMOG grade indicated that this corpus requires only a high school reading level, even lower than that of Reuter news, while the Dale-Chall grade indicated that it needs at least college level reading skill to understand, meaning the EC text is as difficult as that in the PubMed corpus. The most likely explanation for this result is that while EC text uses many short terms, they are unfamiliar to the public, but this cannot be accounted for when using SMOG methods, because SMOG only takes the number of polysyllables into consideration. For
example, some technical terms, e.g. “heparin”, are much shorter than some familiar words, e.g. “characteristics”, but are more difficult to understand for lay readers. Moreover, we also observed a large amount of professional abbreviations present in EC texts (e.g. “AD”, “KPS”), which also have fewer syllables, but present more difficulties for readers with low health literacy. The Dale-Chall scoring system, in contrast to SMOG, takes familiarity of the words into major consideration, perhaps explaining why it determined EC texts to be as difficult to read as scientific papers. Thus, in conclusion, for EC texts, the syntactic structure may not cause problems for readers, but the large number of unfamiliar terms (e.g. technical jargon) used can be a major obstacle for lay readers, who lack a high level of health literacy. The other parts of the text in clinical trials literature were found to require an even higher reading level (see Figure 3). Each dot represents a clinical trial. Most of the dots are above the diagonals, consistent with the recent results for the readability assessment of the detailed descriptions in CT texts from clinicaltrials.gov by Wu et al. Specifically, in their paper, they reported that detailed descriptions required an education level of grade 18 to understand. Therefore, it seems that other parts of the CT texts cannot offer much help to readers with respect to understanding eligibility criteria.

Figure 3. Lexical feature comparison between eligibility criteria and the according whole documents. Left: SMOG Grade; Right: Dale-Chall Grade. The x-axis is the SMOG grade of the whole clinical trial summary and the y-axis is the SMOT grade of the eligibility criteria text.

2. Term familiarity evaluation

The results for lexical features in Table 3 have led us to conclude that the terms used in EC texts are as unfamiliar for the public as those used in scientific papers, even though sometimes the words or terms are not that long and complex. Those findings agree with our hypothesis that technical jargon is one of the major obstacles to lay readers comprehending EC texts. To further confirm this hypothesis, we performed a term familiarity evaluation. As described in the methods section, we chose 90 UMLS concepts to evaluate for term familiarity by computing the KL divergence of the EC texts from both the PubMed and the Reuter corpora. If the divergence of term usage between EC texts and PubMed articles is smaller than that of the news articles, then the conclusion would be that EC are organized and composed with unfamiliar and more technical terms like professional papers. Otherwise, the reverse would be true. For instance, CUI C0027051 has the lexical variants “heart attack” and “myocardial infarction.” In our eligibility criteria text, “heart attack” occurred 434 times while “myocardial infarction” occurred 5810 times, accounting for 7% and 93% of the text, respectively. In Reuter, this set of frequencies was 100% and 0%; while in the PubMed corpora, it was 50% and 50%. The results for KL divergence showed that for this set of lexical variants, the eligibility criteria text was closest to PubMed articles. The 90 comparison results after 180 times performing KL divergence are shown in Figure 4. Only eight sets of lexical variants from the same CUI could not determine which

Figure 4 Results of KL divergence for 90 chosen UMLS CUIs to perform word familiarity prediction.
corpus our eligibility criteria was closer to. Among the rest of the valid values obtained, about 60 % of the concept usage results indicated a more technical lexical variant of the same concept was likely to be chosen to elaborate the EC. Again, this kind of term usage pattern is likely to increase the difficulty for out-of-domain readers to understand the criteria.

3. Clinical Trial Stratification analysis

The readability assessment of the EC texts from clinicaltrials.gov has come to a conclusion, but in order to make use of our results to improve the readability of clinical trials, we extended our research one step further and did a series of preliminary analyses to guide future work and in the hopes that we could find factors related to the changes in required reading level. We stratified the trials according to different features: source country, study type, recruitment conditions and whether they had final results. From all 120,977 files used for general analysis, we filtered out the data that had missing records for these features, which left 46,137 trial files to be examined in our stratified analysis.

Start year and study results.

Only 4,449 of the 46,137 files, less than 1/10 of all the clinical trials examined, included final study results. This is depicted in Figure 5 – where one dot represents one documented trial. There are several possibilities that may lead to no result being recorded. Often it’s either because the trial was not completed yet or the studies were terminated halfway. As the graphs also show, the red dots (trials with recorded results) are all located on the right side, indicating that the clinical trials with recorded results are more likely to be the trials started more recently. One way to explain this fact could be that the related knowledge and findings have grown tremendously in recent years, or the technology to document related records has advanced. For this reason, we were not surprised to observe a significant increase in numbers of records in the early 21st century, with the variance in required reading level of the eligibility

![Figure 5](image.png)

**Figure 5.** Distribution of lexical features among years according two readability measures. The red dots present the trials with final results, and the green points represent this without results records criteria being enlarged since then as well.

Study type.

The two graphs in the upper row describe the readability of lexical features for eligibility criteria, and show no significant distinction among the different types of trials. However, when evaluating the whole documents (CT texts; two graphs in the lower row of **Figure 5**), the red and purple dots clearly split into separate groups, with the group of purple dots, representing interventional trials, clearly requiring above-average literacy to comprehend. This separation indicates that the discrepancy in required reading levels between the two major kinds of trials is more distinct than when considering EC texts alone. The results also show that, the observational studies were much easier to read than the interventional studies, which is consistent with the internal complexity of the interventional studies. However, as the number of submitted clinical trials began to increase rapidly, the readability of the two types converged to be more similar.

This very interesting phenomenon requires further study to determine what factors influenced this convergence. In comparing the results for EC texts alone (upper row in Fig 5) and the whole trial documents (lower row), it’s obvious that the complexity of the interventional studies with respect to reading is not mainly caused by the eligibility criteria text but by other parts of the trial documents, e.g., detailed descriptions of the studies. Figure 6
reveals that observational (purple) and interventional (green) studies comprise the majority of the documents. It also shows that, as the 21st century approached, the number of interventional studies increased dramatically, a reasonable result given the advances made in medical science.

Source country.

Before the study, we hypothesized that whether English was the first language for the researchers who wrote those trial summaries might also influence the readability of the trial documentations. Since the number of U.S. and Non-U.S studies in the corpus were relatively equivalent (39% and 45%), we were able to classify trials for source
country into one of only two groups: U.S. or non-U.S. (includes many non-English speaking countries, e.g. Germany, China, Japan). We then explored the differences in the readability of the trials according to country where the trial was being performed. It’s interesting to notice that when stratified the whole documents (the two graphs in lower row of Figure 7), the trials with no location recorded showed either much higher or lower readability than the average level. We sorted all the trials by their startyears and noticed that most of the trials short of country information were started before 2000, and account for a large part of all documents before 1990s. We manually reviewed some trial documents with extrem readability scores and noticed that the low scores (easy to read) were always gained because the incompleteness of the trial documents, while the extremely high scores (difficult to read) were caused by multiple reasons, e.g. internal complexity of the study disease. One intersting thing we also found is that most of the low-readability trials without country information while started in recent year (after 2000) were those sponsored by large international pharmaceutical companies (which cause no country information recorded), and this kind of trials tend to be more difficult for readers compared to those conducted by federal-funded research institutes, universities or hospitals.

**Discussion**

**Implications of the findings**

According to our results, to gain a full comprehension of contents of EC, the readers have to be at least college education level, which is higher than the average American, who has achieved a high-school level. Therefore, in order to allow the many kinds of people who are looking for information in this online database, especially out-of-domain readers, to understand EC, there is a need for the website and researchers to be more careful in expressions and take those health consumers who are at a lower education level into consideration. Particularly, it was found that the main difficulty in understanding the contents is not a result of the syntactic features, but of the lexical ones, meaning that it is not a lot of long, complex sentences confusing people, but rather, the technical terms frequently used that increases the reading level required for target readers. Hence, the readability problem might be solved by using fewer technical jargons including clinical terminology and abbreviations, and instead considering lexical variants designed for patients, like consumer health vocabularies.

To realize this goal, we need to establish a comprehensive consumer health dictionary. However at present, no existing consumer health terminology can be called comprehensive or complete. It is a non-trivial effort to develop a comprehensive terminology that can keep up with the face pace of the constantly evolving consumer vocabulary. Most of the existing consumer vocabulary lists are based in a particular disease domain or user community. Given this situation, using existing consumer terminology to translate from clinical terminology is not realistic. Therefore, for further development, either existing “partial” consumer health vocabulary lists could be used to translate clinical trials in different fields, which lists could then be linked together, or an attempt could be made to find a method to establish a new comprehensive consumer health terminology, which would be a huge, long-term research effort.

**Limitations**

This study has several limitations. First, we only used computational measures to estimate readability. In the future, it would be valuable to engage enough real users, including both patients and clinicians, to validate our computational results and to obtain their feedback on their real-life experience. Second, the stratified analysis only serves as preliminary work for future guidance. The interpretation of the stratified analysis results here is just some reasonable guessing and is one of many possibilities without solid statistical tests to confirm. In the future, more relative study should be carried out based on refined stratification to find more interactive information.

**Conclusions**

In this study we presented an initial assessment of the readability of clinical trials eligibility criteria in ClinicalTrials.gov. We used both general popular readability formulas and health-domain specific methods, covering three aspects of readability: the unit length of a text, the syntactic complexity, and the lexical complexity. Moreover, we also conducted a comparison with two other health-related corpora to assess the term use tendency of the clinical trials. The overall results showed that eligibility criteria texts are beyond the comprehension capacity of the general American population, and the main reason for that is the frequent use of professional terminology and technical jargons. To guide future works on improvement, we also classified the trial texts by different factors and study the readability, trying to find some properties of the trials that might impact the readability. In the results, we found that in the beginning of the 21st century, the number of documented clinical trials increased greatly. Of all the properties we have studied, different study types of clinical trials might result in different levels of readability: generally, interventional studies require a higher education level to understand than observational studies. We also found that
clinical trials sponsored by pharmaceutical companies tend to have higher reading level requirement compared to federally-funded studies. Overall, our study presented a systematic evaluation and analysis of this online clinical trials database and contributed evidence for future improvement.

Acknowledgments

This study was sponsored by the U.S. National Library of Medicine grant R01LM009886 (PI: Weng) and U.S. National Center for Advancing Translational Science grant UL1 TR000040 (PI: Ginsberg).

References

Automatic Assignment of Non-Leaf MeSH Terms to Biomedical Articles

Ramakanth Kavuluru, Ph.D1, 2 and Anthony Rios, B.S2

1Division of Biomedical Informatics, Department of Biostatistics, University of Kentucky  
2Department of Computer Science, University of Kentucky

Abstract

Assigning labels from a hierarchical vocabulary is a well known special case of multi-label classification, often modeled to maximize micro $F_1$-score. However, building accurate binary classifiers for poorly performing labels in the hierarchy can improve both micro and macro $F_1$-scores. In this paper, we propose and evaluate classification strategies involving descendant node instances to build better binary classifiers for non-leaf labels with the use-case of assigning Medical Subject Headings (MeSH) to biomedical articles. Librarians at the National Library of Medicine tag each biomedical article to be indexed by their PubMed information system with terms from the MeSH terminology, a biomedical conceptual hierarchy with over 27,000 terms. Human indexers look at each article’s full text to assign a set of most suitable MeSH terms for indexing it. Several recent automated attempts focused on using the article title and abstract text to identify MeSH terms for the corresponding article. Despite these attempts, it is observed that assigning MeSH terms corresponding to certain non-leaf nodes of the MeSH hierarchy is particularly challenging. Non-leaf nodes are very important as they constitute one third of the total number of MeSH terms. Here, we demonstrate the effectiveness of exploiting training examples of descendant terms of non-leaf nodes in improving the performance of conventional classifiers for the corresponding non-leaf MeSH terms. Specifically, we focus on reducing the false positives (FPs) caused due to descendant instances in traditional classifiers. Our methods are able to achieve a relative improvement of 7.5% in macro-$F_1$ score while also increasing the micro-$F_1$ score by 1.6% for a set of 500 non-leaf terms in the MeSH hierarchy. These results strongly indicate the critical role of incorporating hierarchical information in MeSH term prediction. To our knowledge, our effort is the first to demonstrate the role of hierarchical information in improving binary classifiers for non-leaf MeSH terms.

1. Introduction

Indexing biomedical articles is an important task that has significant impact on how researchers search and retrieve relevant information. This is especially essential given the exponential growth of biomedical articles indexed by PubMed®, the main search system developed by the National Center for Biotechnology Information (NCBI). PubMed lets users search over 22 million biomedical citations available in the MEDLINE bibliographic database curated by the National Library of Medicine (NLM) from over 5000 leading biomedical journals in the world. To keep up with the explosion of information on various topics, users depend on search tasks involving Medical Subject Headings (MeSH®) that are assigned to each biomedical article. MeSH is a controlled hierarchical vocabulary created by the NLM and consists of medical subjects that form a directed acyclic graph (DAG). Once articles are indexed with MeSH terms, users can quickly search for articles that pertain to a specific subject of interest instead of relying solely on keyword based searches.

Since MeSH terms are assigned by librarians who look at the full text of an article, they capture the semantic content of an article that cannot easily be captured by keyword or phrase searches. Thus assigning MeSH terms to articles is a routine task for the indexing staff at NLM. This is empirically shown to be a complex task with 48% consistency because it heavily relies on indexers’ understanding of the article and their familiarity with the MeSH vocabulary [1]. As such, the manual indexing task takes a significant amount of time leading to delays in the availability of indexed articles. It is observed that it takes about 90 days to complete 75% of the citation assignment for new articles [2]. Moreover, manual indexing is also a fiscally expensive initiative [3]. Due to these reasons, there have been many recent efforts to develop automatic ways of assigning MeSH terms for indexing biomedical articles including an on-going indexing challenge (http://www.bioasq.org/). However, automated efforts (including ours) mostly focused on predicting MeSH terms for indexing based solely on the abstract and title text of the articles. This is because most full text articles are only available based on paid licenses not subscribed by many researchers.

Many efforts in MeSH term prediction generally rely on two different methods. The first method is the $k$-nearest neighbor ($k$-NN) approach where $k$ articles that are already tagged with MeSH terms and whose content is found to be “close” to the new abstract to be indexed are obtained. The MeSH terms from these $k$ articles constitute candidate terms for the new abstract [2]. A second method is based on applying machine learning algorithms to learn and
index the best binary classifier models for each MeSH term. A new candidate abstract would then be put through all these classifiers and the corresponding MeSH terms of classifiers that return a positive prediction are chosen as candidate terms for the abstract. This approach has been termed as meta-learning and researchers at the NLM [4] have demonstrated that depending on the individual terms different classifiers might yield different results thus justifying the approach.

In one of their recent results, Jimeno-Yepes et al. [5] identify six MeSH terms that have been found to be difficult to predict using existing indexing approaches at the NLM. Out of these, two terms, hormones and infection, were shown to have unusually high numbers of false positive instances (with existing classification approaches) that are actually tagged with the descendants of the corresponding terms in the MeSH hierarchy. That is, there were many citations (abstract and title text) that were originally assigned a more specific descendant of an ancestor term but where the ancestor term classifier has classified each of them as a positive instance. Using this as motivation, in this paper, we explore classification methods that utilize training examples from the descendants as a way of improving non-leaf MeSH term prediction.

The rest of the paper is organized as follows. In the rest of this section, we first discuss the essential background on the size and structure of the MeSH terminology and discuss some related efforts. We also discuss a motivating scenario for incorporating descendants of non-leaf terms in building non-leaf term binary classifiers. In Section 2, we discuss our main methods that exploit descendant training instances to reduce false positive errors. We present our experiments using a large dataset and the corresponding results obtained with associated discussion in Section 3.

1.1. Background and Related Work

MeSH’s main purpose is to index biomedical articles and hence strict notions of meronymy were not used in its design; the hierarchical relationships are actually guided by “aboutness” of a child to its parent\(^1\). Hence a term could be a descendant of multiple other terms whose least common consumer is not one of them. That is, a term could have multiple paths from the root. In the 2013 version of MeSH, there are 26,578 main subject headings. The scatter plot of number of descendants and depth of the term is shown in Figure 1. Note that there are 16 individual hierarchies in MeSH but the machine processable UMLS Metathesaurus\(^2\) files combine these hierarchies with additional placeholder nodes to have a unified root. In the figure, the depth of the root is taken as 0 and the first valid MeSH nodes for each of the 16 hierarchies start at depth 1. From the figure, we can see that the number of descendants is in hundreds for many nodes at depths \(\leq 4\). By analyzing the MeSH hierarchy (Figure 2), we also found that there are at least 15,000 MeSH nodes with depth \(\leq 4\).

![Figure 1: Scatter Plot of Term Depth and Number of Descendants](image)

NLM initiated efforts in automatic MeSH term extraction with their Medical Text Indexer (MTI) program that uses a combination of \(k\)-NN based approach and named entity recognition (NER) based approaches with other unsupervised clustering and ranking heuristics in a pipeline [6]. MTI recommends MeSH terms for NLM indexers to assist in their efforts to expedite the indexing process\(^3\). Another recent approach by Huang et al. [2] uses \(k\)-NN

---

\(^1\)http://www.nlm.nih.gov/mesh/meshrels.html
\(^2\)http://www.nlm.nih.gov/pubs/Factsheets/umlsmeta.html
\(^3\)For the full architecture of MTI’s processing flow, please see: http://skr.nlm.nih.gov/resource/Medical_Text_Indexer_Processing_Flow.pdf
approach to obtain MeSH terms from a set of $k$ already tagged abstracts and use the learning to rank approach to carefully rank the MeSH terms. They use two different gold standard datasets one with 200 abstracts and the other with 1000 abstracts and achieve an F-score of 0.5 and recall 0.7 on the smaller dataset compared to MTI’s F-score of 0.4 and recall 0.57. Vasuki and Cohen [7] also use the $k$-NN approach but employ reflective random indexing to find the nearest neighbors in the training dataset and use the indexing based similarity scores to rank the terms from the neighboring citations.

Several other attempts incorporated different machine learning approaches with novel feature selection [8] and training data sampling [9] techniques. In our earlier effort, we also explored purely unsupervised approaches that rely on term co-occurrence counts and named entity recognition [10], which we subsequently extended [11] to a supervised framework exploiting latent associations between MeSH headings based on reflective random indexing. A recent effort by Jimeno-Yepes et al. [4] uses a large dataset and uses meta-learning to train custom binary classifiers for each MeSH term and indexes the best performing model for each label to be applied on new abstracts; we request the reader to refer to their work for a recent review of machine learning methods used for MeSH term assignment. In a related effort [5] they identify six MeSH terms that have been found to be difficult to predict using existing indexing approaches at the NLM. In this work, we look at all non-leaf nodes ($8756 \approx 33\%$ of all MeSH terms) in the MeSH hierarchy with varying number of descendants For these terms we develop and evaluate binary classifiers that utilize examples of their descendants. Given recent efforts [4, 12] that heavily rely on accurate binary classifiers for each term, we believe our work is very relevant to further the state-of-the-art in automated indexing of biomedical articles.

1.2. Motivating Scenario & Baseline Approach

In this section we demonstrate the presence of false positives arising from descendant nodes using the following six terms with varying descendant counts.

1. **Membrane Proteins**: 971 descendants
2. **Neoplasms**: 663 descendants
3. **Hormones**: 221 descendants
4. **Infection**: 154 descendants
5. **Mutation**: 47 descendants
6. **Plasmodium**: 10 descendants

Here *Hormones* and *infection* are from the challenging terms used in experiments in [5]; among the terms considered in these related efforts, these two had many descendants and high descendant attributable FPs. In addition to these, we chose four other terms arbitrarily with varying numbers of descendants. For our motivating scenario, we built training and testing datasets for each of these terms by extracting positive and negative examples from over 20 million Medline citations that are already fully tagged with MeSH terms by trained coders at the NLM. We made sure that our training and testing examples both had a non-empty abstract field as books and other artifacts indexed might not have an abstract. The number of positive and negative examples for each of the terms in the training and testing data are shown in Table 1.

In Table 1, the sum of positive examples in both training and testing sets for each term is equal to the total number of citations (with a non-empty abstract field) tagged with that term in the Medline citation database (2014 Baseline).
We randomly selected approximately 5% of the available citations for each term to be included as the corresponding positive examples in the testing data. Most MeSH terms are tagged for very few citations and hence, we chose a very high number of negative examples as shown in the negative testing instance count column; the positive examples constitute a maximum of around 2% of the total testing dataset for each term.

<table>
<thead>
<tr>
<th>MeSH Term</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Membrane Proteins</td>
<td>98401</td>
<td>340780</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>134082</td>
<td>338308</td>
</tr>
<tr>
<td>Hormones</td>
<td>15679</td>
<td>340416</td>
</tr>
<tr>
<td>Infection</td>
<td>12930</td>
<td>340327</td>
</tr>
<tr>
<td>Mutation</td>
<td>230905</td>
<td>336020</td>
</tr>
<tr>
<td>Plasmodium</td>
<td>3206</td>
<td>343440</td>
</tr>
</tbody>
</table>

For our experiments, we chose as baseline the LIBLINEAR [13] logistic regression implementation through the scikit-learn [14] framework. We used unigrams and bigrams, with a minimum frequency of five citations per n-gram, as our binary features. After applying the classifiers on the corresponding testing sets for each of the six terms, we looked at the number of FPs that can be attributed to descendants of the six terms under consideration. These observations are shown in Table 2 from which we notice that a significant number of FPs (more than 40% for four terms) are due to misclassification of descendants. Based on these observations, we set out to design and evaluate classification approaches that exploit descendant instances.

<table>
<thead>
<tr>
<th>MeSH Term</th>
<th># FPs</th>
<th># Desc. FPs</th>
<th>% Desc. FPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membrane Proteins</td>
<td>7547</td>
<td>3126</td>
<td>41%</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>8133</td>
<td>3533</td>
<td>44%</td>
</tr>
<tr>
<td>Hormones</td>
<td>1349</td>
<td>790</td>
<td>56%</td>
</tr>
<tr>
<td>Infection</td>
<td>1283</td>
<td>356</td>
<td>28%</td>
</tr>
<tr>
<td>Mutation</td>
<td>13268</td>
<td>1733</td>
<td>13%</td>
</tr>
<tr>
<td>Plasmodium</td>
<td>227</td>
<td>157</td>
<td>69%</td>
</tr>
</tbody>
</table>

2. Methods: Hierarchical Strategies to Handle Descendant FPs

Motivated by observations in Table 2, we proceed to exploit instances of descendants terms in building better binary classifiers for non-leaf MeSH terms. Here, the descendant set of a given MeSH term is defined recursively as the union of the set of its children and all their descendants in the MeSH hierarchy. The baseline method is simply building binary classifiers with unigram and bigram features using support vector machines (SVMs). Next, we outline the five different methods that incorporate descendant instances to be compared with the baseline approach. We eventually use a dataset of one million biomedical citations and all non-leaf terms that are assigned to at least 100 citations in this dataset. Details of dataset and results achieved using our methods will be elaborated in Section 3.

Before we proceed, we first note that it is a valid indexing approach to assign a term and one or more of its descendants to the same biomedical article. NLM’s multi-label assignment process allows this when the content of the article warrants such an assignment. As such, the task of assigning MeSH terms is a relaxation of the non-mandatory leaf node prediction [15, Section 4.4] scenario encountered in hierarchical classification. Given this, to best differentiate a term’s instances from its descendants’ instances, we curate training data of ancestor terms that are not tagged with any of their descendants. Similarly, we curate training instances of descendant terms that are not tagged with the corresponding ancestors. If \( T \) is the size of the training set of a given ancestor term selected as explained above, we randomly pick \( T/d \) training instances for each of its \( d \) descendant terms to approximately create equal
sized datasets. While we allow appropriate data imbalance to persist in our main experiments, we selected balanced datasets to distinguish between main terms and their descendants, with equal contribution from each descendant. This is due to the presence of extremely large number of instances when summed over all descendants of a given term. Because of the way we incorporate these descendant instances into our methods (the rest of the section), we believe this descendant specific instance selection is appropriate given the testing data set is chosen randomly while preserving realistic imbalance. Next, we describe each of the five methods that employ descendant training instances.

2.1. **Multiclass Prediction with a Third Descendant Class**

Unlike the binary classification approach used in our baseline method, in this approach, for each term we model the prediction as a multiclass text classification problem with three classes: 1. main non-leaf term class of interest to us, 2. its descendants class, and finally 3. the ‘other’ class that has negative examples of all other non-descendant instances. The non-descendant ‘other’ instances are citations in our dataset that are not tagged with the non-leaf term of interest or its descendants. The positive class instances are those that are tagged with the non-leaf term but none of its descendants. The descendants’ training instances are curated as explained in the beginning of this section with equal contributions from each descendant. We used the built-in one-vs-all approach in LIBLINEAR with the same set of features used for our baseline method in Section 1.2.

2.2. **Two Stage Hierarchical Prediction**

Here we follow the conventional hierarchical approach and build two different classifiers. The top level binary classifier combines instances of either the main term or any of the descendants as one positive class, the negative class being the ‘other’ class from Section 2.1. Intuitively, this classifier is expected to identify testing instances of a candidate term or its descendants. Once an instance is predicted as positive with this classifier, a second binary classifier that distinguishes between the main term and its descendants’ instances is applied. This second classifier is built using the main non-leaf term class and the descendants class from Section 2.1.

2.3. **Baseline Result Filtering using Descendant Classifier**

Instead of applying the second stage classifier in Section 2.2 to the positive instances of the top level classifier in the hierarchical approach, we simply use it to filter the positive instances output by the baseline classifier from Section 1.2. Given the high proportion of FPs attributable to descendant instances (Table 2) and the reasonably high performance we observed in the 2nd stage classifier in the two stage hierarchical approach, we chose to do this filtering.

2.4. **Baseline Filtering with Descendant Classifier and FN recovery**

Our primary focus thus far has been on minimizing FPs. In this context, clearly, the percentages of total FP reduction should be looked at in the context of the recall loss due to new false negatives (FNs) induced. Just as many FPs can be attributed to descendant instances, it is also possible that several FNs can be attributed to descendants. This is because of the multi-label indexing nature where both a term and any of its descendants can be assigned to a given article as discussed in the beginning of this section. That is, many negative predictions that end up in the descendant class in Section 2.1 or of the second stage classifier in Section 2.3 could actually be positives. To reduce these descendant attributable FNs, we propose to move several of these positives from the incorrect descendant predictions to their correct ancestor predictions using a binary FN recovery classifier that distinguishes between instances that are tagged with “only descendants” (negative class) and instances that are tagged with “both the ancestor term and at least one descendant” (positive class). We apply this binary classifier to the negative predictions from the second stage classifier of the baseline filter approach (Section 2.3). Those that are classified as belonging to the class “both the ancestor term and at least one descendant” are considered positive instances of the non-leaf ancestor term in the end. To build this binary classifier, we obtain the positive instances by considering all citations tagged with the non-leaf term and any of its descendants. The negative instances dataset is again a subset of the descendant-only examples as explained at the beginning of this section.

2.5. **Combining NER and Supervised Classification**

Our final approach involves combining named entity recognition with supervised classification. It is also essentially a filter on positive classifications of the baseline classifier from Section 1.2. In this method, we first use a state-of-the-art biomedical named entity recognition tool, MetaMap [16], on the input citation of each positive instance output.
by the baseline classifier to identify MeSH terms mentioned in its title and abstract. If any of these MeSH terms is actually a descendant of the candidate term, we subject it to a corresponding new binary classifier that distinguishes between instances that are tagged with only descendants and those tagged with both a main term and at least one of its descendants (from Section 2.4). Those that are classified into the “only descendants” are treated as negatives in this filter. Intuitively, since the input instances to this filtering approach are already deemed positive for the main term (since they are the positives from the baseline classifier), we only need to identify those instances that should be exclusively tagged with descendant terms only. However, just because a citation contains a descendant MeSH term does not necessarily mean it can only take a descendant MeSH term as a tag, although it could be a candidate for further analysis. Hence we apply the supervised classifier to those that contain a descendant term.

2.6. Summary of Our Approaches

Some of our methods in Sections 2.1–2.5 involve multiple classifiers that use different training data subsets. For clarity, here we summarize all the classifiers involved and how they are used in different methods using the corresponding section numbers. We use the following notation to represent different training data subsets for our classifiers.

- **A**: set of all training examples annotated with the main non-leaf term but none of its descendants
- **D**: set of examples not tagged with the main non-leaf term but with at least one of its descendants
- **AD**: examples tagged with both the non-leaf term and at least one of its descendants
- **O**: training examples that are not in sets A, D, or AD.

### Table 3: Classifiers Used in Different Methods

<table>
<thead>
<tr>
<th>(a) Classifier-Dataset Map</th>
<th>(b) Method-Classifier Map</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset</strong></td>
<td>L1</td>
</tr>
<tr>
<td>A</td>
<td>+</td>
</tr>
<tr>
<td>D</td>
<td>-</td>
</tr>
<tr>
<td>AD</td>
<td>+</td>
</tr>
<tr>
<td>O</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on this notation, in Table 3(a), we identify the five different types of classifiers L1–L5 used in our methods and the corresponding training data subsets used. All classifiers are binary except L2, which is a three-way classifier. For L4 and L5, the D set is sampled as elaborated in the second paragraph of Section 2. In Table 3(b), we specify the classifiers used for the baseline approach (Section 1.2) and our proposed FP handling methods in Sections 2.1–2.5.

3. Results and Discussion

To ensure a natural distribution of terms, we chose a set of 1 million already annotated citations with dates of publication between 2005–2014 from around 1900 well known journals used in the BioASQ competition to index biomedical articles. Out of this dataset, 850,000 citations were used for training, 100,000 were used for validation, and the remaining 50,000 were used for testing. The testing dataset instances’ dates of publication are chronologically later than the training and validation dataset instances. Among all non-leaf terms in the MeSH hierarchy, we chose those that had at least 100 instances in the training dataset, which resulted in 4395 non-leaf terms that we considered for our experiments. Before we proceed we outline the performance measures used.

3.1. Performance Measures

Since the task of assigning multiple MeSH terms to a citation is the multi-label classification problem, there are multiple complementary methods [17] for evaluating automatic approaches for this task. Here, we focus on label based measures since we are only concerned about the performance of non-leaf MeSH terms.
For each MeSH heading $T_j$ in the set of terms $T$ being considered, we have label-based precision $P(T_j)$, recall $R(T_j)$, and $F_1$-score $F(T_j)$ defined as

$$P(T_j) = \frac{TP_j}{TP_j + FP_j}, \quad R(T_j) = \frac{TP_j}{TP_j + FN_j}, \quad \text{and} \quad F(T_j) = \frac{2P(T_j)R(T_j)}{P(T_j) + R(T_j)}$$

where $TP_j$, $FP_j$, and $FN_j$ are true positives, false positives, and false negatives, respectively, of term $T_j$. Given this, the label-based macro average $F_1$-score is

$$\text{Macro-F} = \frac{1}{|T|} \sum_{j=1}^{|T|} F(T_j).$$

The label-based micro precision, recall, and $F_1$-score are defined as

$$P_{\text{mic}} = \frac{\sum_{j=1}^{|T|} TP_j}{\sum_{j=1}^{|T|} (TP_j + FP_j)}, \quad R_{\text{mic}} = \frac{\sum_{j=1}^{|T|} TP_j}{\sum_{j=1}^{|T|} (TP_j + FN_j)}, \quad \text{and} \quad \text{Micro-F} = \frac{2P_{\text{mic}} \cdot R_{\text{mic}}}{P_{\text{mic}} + R_{\text{mic}}}.$$  

While macro measures consider all labels as equally important, micro measures tend to give more importance to more frequent labels. Here we are interested in improving both the micro and macro averages over non-leaf terms.

### 3.2. Application of Hierarchical Classifiers

We first applied our hierarchical strategies (Section 2) to the set of six terms in the motivating scenario in Section 1.2. All approaches showed some improvement except the two-stage hierarchical approach in Section 2.2, which suffered a major loss in precision. This is not surprising given recent attempts by other researchers [18] also show that top-down hierarchical approaches may not be very suitable with very large terminologies with thousands of labels. Hence all our large scale experiments described in this section are conducted with all the methods except the two-stage approach. We note that all approaches we consider here are thus essentially filters on the positive instances of the baseline. That is, we are not attempting to obtain better recall than baseline but are willing to trade off some recall for precision increases that could yield us better micro and macro $F$-measures.

Instead of using the default approach of making predictions based on SVM scores, we use a more suitable approach for multi-label classification based on the meta-labeler [19] method where, in addition to a ranked list of labels (based on probability estimates) for each artifact to be classified, a threshold on the number of labels is also predicted using the same set of features used for predicting the labels. This generally helps because some binary classifiers for labels with low frequencies do not always produce high scores for the positive class. We used this approach and built over 27,000 binary classifiers (one for each MeSH term even if it is not a non-leaf term) using the training dataset. We also developed another model that predicts the number of terms for an input citation based on the training dataset. Next, we applied each of these classifiers to each instance in the validation dataset. We ranked the MeSH terms for each instance based on the corresponding SVM classifier scores and chose the threshold for the number of terms to retain based on the thresholding model. Next, for each of our hierarchical strategies, if one of the 4,395 non-leaf terms that we consider in our experiments shows up in the top terms predicted using the baseline approach and the thresholding, we apply the hierarchical strategy as an additional filter. If the output of the hierarchical strategy also ends up in the positive class, we take it as a positive classification for the non-leaf term. If the strategy outputs a negative prediction, we treat the prediction to be negative for the non-leaf term. Thus, we reverse some positive non-leaf term predictions by the base classifier using the hierarchical strategies.

### 3.3. Results on Validation Dataset

We used the validation dataset to identify non-leaf MeSH terms for which we notice an improvement over the baseline when the hierarchical strategies are used. We then applied the strategies for only such non-leaf terms on the test dataset. This is to ensure that our choice of the terms is not influenced by the test data. Here we present our results on the validation dataset. In Table 4, for each strategy (first column, indicated by the section number), we show how many terms had better $F_1$-score (compared with the baseline), the average increase in macro-$F_1$ and micro-$F_1$ over such terms, and average number of descendants and average depth for such terms. The final row is a hybrid classifier that chooses the best hierarchical strategy over the baseline for each non-leaf term.

From the table, we see that the FP filtering approach with a descendant classifier and an FN recovery classifier (Sec 2.4) had the largest improvement when considering both macro and micro averages. Although the NER based approach (Sec 2.5) improved results for over 300 terms, the improvement is smaller compared to other methods. The average depth of the full MeSH hierarchy is 4.5 with an average of 17 descendants per non-leaf term. From Table 4, it is evident that the hierarchical strategies seem to do well when the terms are slightly higher in the hierarchy with
many descendants. The absolute improvement is around 1.5% in both micro and macro averages for a total of 536 terms, which constitute about 12% of the set of non-leaf terms we considered. We will now apply the hierarchical approaches as discussed in Section 3.2 to the test dataset.

### 3.4. Results on Test Dataset

The test set results in Table 5 are obtained by considering all those non-leaf terms that had at least 1% improvement in $F_1$ score in the validation dataset and ranking them in the descending order of their frequency in the validation dataset. This is to analyze if we can find a large set of terms for which the performance holds across different datasets. From the table it is clear that both micro and macro averages increase for the top 500 non-leaf terms. We also display relative improvements ($\frac{(\text{hybrid} - \text{baseline})}{\text{baseline}}$) in the table which become prominent at the macro level as we consider fewer frequently occurring terms. But overall were able to identify 500 terms for which we improved micro-$F_1$ score by 1.6% and macro-$F_1$ score by 7.5% relative to the baseline scores. These results indicate the strong potential of exploring hierarchical information in improving the state-of-the-art in automated indexing of biomedical articles.

#### Table 5: Test Dataset Results

<table>
<thead>
<tr>
<th></th>
<th>Micro-$F$</th>
<th>Macro-$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Hybrid</td>
</tr>
<tr>
<td><strong>Top 100</strong></td>
<td>0.6304</td>
<td><strong>0.6421</strong></td>
</tr>
<tr>
<td><strong>Top 250</strong></td>
<td>0.6106</td>
<td><strong>0.6201</strong></td>
</tr>
<tr>
<td><strong>Top 500</strong></td>
<td>0.5959</td>
<td><strong>0.6034</strong></td>
</tr>
</tbody>
</table>

### 3.5. Remarks on Class Imbalance

As noted in Section 1.2, there is extreme imbalance between the positive and negative instances of MeSH terms, a situation which is more prominent in non-leaf terms. We conducted additional experiments with different class weighting schemes for all the strategies discussed in Section 2. To handle high negative class bias in a methodical way we also tried a combination of boosting and random under sampling [20]. Although there were marginal improvements for a few terms, we do not think they warrant reporting in this manuscript and so we leave them out. However, this also leads us to believe that this problem is not only important but also very difficult to solve, especially under the assumption that only the abstract and title are available. There is some evidence [21] that points to small improvements if additional sections of the article full text, when available, are considered. The class imbalance in the MeSH training data can also result from the so called “Rule of Three”\(^5\) for indexing biomedical articles – “If more than 3 related concepts are discussed in an article but are not a major topic, the more general MeSH heading under which they are all treed is usually indexed. The specific headings usually are not indexed." Due to this rule, some of the false positives could have been due to more than three specific MeSH headings predicted for a citation with a common immediate parent. So we conducted additional experiments where we replaced such term sets with their immediate ancestor in the final candidate set of predicted terms (an extension to the method in Section 3.2).

However, we observed only marginal improvements; incorporating this particular manual indexing guidance into the automation process did not seem to help. The rule-of-three also states that when three or more related concepts are identified as major MeSH terms (terms that are deemed central to the article), their immediate common ancestor is coded as a major topic, while the descendants are retained but not as major terms. Thus this aspect of the rule-of-three is also essentially responsible for situations where both a non-leaf term and several of its descendants are assigned to an article.

For our initial study we used a dataset of one million citations although there are currently about 21 million citations made available through NLM for training purposes. Although we chose the smaller dataset for computational tractability in building over 27,000 baseline models and four sets of 4,395 hierarchical models for non-leaf terms, increasing the dataset size to 22 million citations may not necessarily yield improved performance. This is because of the extreme class imbalance that will persist given the unusually high number of negative examples for each non-leaf term. In such cases, it is not clear whether rare event adjustments to supervised learning approaches [22, 23] will help the situation for MeSH term prediction because of the limitation of considering only the title and abstract. However, there is recent evidence that micro averages (over all MeSH terms) can be improved by using more efficient learning methods and considering all available citations [24]. Our main purpose in this paper is to show that descendant instances play an important role in improving results for many MeSH terms. We believe their role will only increase in relevance, if extreme class imbalances are handled, because of the hierarchical nature of MeSH.

4. Conclusion

Assigning MeSH headings to biomedical articles is an important task at the NLM due to the search flexibility it gives biomedical researchers in satisfying their information needs. Many automated attempts have been developed in the recent past to predict MeSH terms from the title and abstract text of an article. In this paper, we focus on the problem of building binary classifiers for terms corresponding to non-leaf nodes in the MeSH hierarchy. Based on the observation that many false positives for terms with several descendants could be attributed to the descendants, we experimented with approaches that exploit descendant training instances. To our knowledge, this is the first effort to investigate the suitability of hierarchical information toward improving MeSH term prediction. Our results show good promise with a relative improvement of 7.5% in macro-\(F_1\) score and 1.6% in micro-\(F_1\) score over the baseline non-hierarchical approach for a set of 500 non-leaf terms that occur at least 100 times in our training dataset of one million citations.

Given that there are over 23 million citations and the recent evidence [24] of computational tractability for learning using all citations, we will conduct additional experiments to identify other non-leaf terms whose performance can be improved using hierarchical approaches. We will also conduct additional error analysis to identify factors that make certain non-leaf terms more suitable for hierarchical classification compared to others. Our current effort in this paper, nevertheless, demonstrates that hybrid hierarchical approaches that exploit descendant instances play a key role in assigning MeSH terms to biomedical articles.

Acknowledgments

We are grateful to anonymous reviewers for their careful assessment and constructive criticism of the manuscript, which helped improve it for the final camera-ready version. This publication was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, US National Institutes of Health (NIH), through Grant UL1TR000117. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References


Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation

Mehmet Kayaalp, MD, PhD, Allen C. Browne, MS, Pamela Sagan, RN, Tyne McGee, BS, Clement J. McDonald, MD
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract
The Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. We have been manually annotating clinical text since 2008 in order to test and evaluate an algorithmic clinical text de-identification tool, NLM Scrubber, which we have been developing in parallel. Although HIPAA provides some guidance about what must be de-identified, translating those guidelines into practice is not as straightforward, especially when one deals with free text. As a result we have changed our manual annotation labels and methods six times. This paper explains why we have made those annotation choices, which have been evolved throughout seven years of practice on this field. The aim of this paper is to start a community discussion towards developing standards for clinical text annotation with the end goal of studying and comparing clinical text de-identification systems more accurately.

1. Introduction
The Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. The Rule indicates 18 pieces of personally identifiable information (PII) that need to be de-identified in order to protect patient privacy. Our particular interest in the Privacy Rule is to use it as our guideline for preventing unintended privacy breaches during the secondary use of patient health information for clinical research. Although it is very clear what each piece of PII is, conceptually, it may not be so straightforward when we try to do so manually in a clinical narrative report. The Privacy Rule has been designed mainly with the structured tabular data in mind. When we attempt doing the same with text, we are faced with a number of difficulties that arise due to the nature of English, or any other natural language.

For example, the Privacy Rule requires de-identification of personal names but does not say anything about personal name initials (e.g., JFK). While U.S. District Courts impose restrictions on the use of personal names of minors and require their names in all hearing transcripts to be de-identified by converting them into personal name initials,1-3 the Office of Civil Rights of the Health of Human Services interprets the Privacy Rule in such a way that equates personal name initials with personal names.4

It is unreasonable to expect from any piece of legislation or set of rules to spell out every imaginable version or combination of entities that could occur in a clinical narrative report or patient record. Our approach is to take the Privacy Rule as a model, interpret its language, understand its aim, and in ambiguous cases, make decisions whether we ought to de-identify those particular pieces of information in order to fully comply with the Privacy Rule to the best of our abilities.

To this end, we have been developing annotation guidelines, which basically are a compendium of examples, extracted from clinical reports, to show what types of text elements and personal identifiers need to be annotated using an evolving set of labels. We started annotating clinical text for de-identification research in 2008, and since then we have revised our set of annotation labels (a.k.a. tag set) six times. As we are preparing this manuscript, we are working on the seventh iteration of our annotation schema and the label set, and will be making it available at the time of this publication.

Although the Privacy Rule seems pretty straightforward at first glance, revising our annotation approaches so many times in the last seven years is indicative of how involved and complex the task is. We don’t believe that publishing the guidelines would suffice by themselves, since the guidelines only tell what needs to be done. In this paper, we try to address not only what we annotate but also why we annotate the way we do. We hope that the rationale behind our guidelines would start a discussion towards standardizing annotation guidelines for clinical text de-identification. Such
standardization would facilitate research and enable us to compare de-identification system performances on an equal footing.

Before describing our annotation methods, we provide a brief background on the process and rationale of manual annotations, discuss personally identifiable information (PII) as sanctioned by the HIPAA Privacy Rule, and provide a short overview of approaches of how various research groups have adopted PII elements into their de-identification systems. We conclude with Results and Discussion sections.

2. Background

Manual annotation of documents is a necessary step in developing automatic de-identification systems. While de-identification systems using a supervised learning approach necessitate a manually annotated training set, all systems require manually annotated documents for evaluation. We use manually annotated documents both for the development and evaluation of NLM-Scrubber. 5-7

Even when semi-automated with software-tools, 8 manual annotation is a labor intensive activity. In the course of the development of NLM-Scrubber we annotated a large sample of clinical reports from the NIH Clinical Center by collecting the reports of 7,571 patients. We eliminated duplicate records by keeping only one record of each type, admission, discharge summary etc. The primary annotators were a nurse and linguist assisted by two student summer interns. We plan to have two summer interns each summer going forward.

These annotators used NLM’s Visual Text Tagging tool, VTT. VTT allows annotators to select and annotate strings of text by swiping the cursor over them and choosing a tag from a pull-down list of annotation labels. The application displays the annotation with a distinctive combination of font type, font color and background color. Tags in VTT can have sub-tags which allow the two dimensional annotation scheme described below. VTT saves the annotations in a stand-off manner leaving the text undisturbed and produces records in a machine readable pure-ASCII format. A screen shot of the VTT interface is shown in Figure 1. VTT has proven helpful both for manual annotation of documents and for displaying machine output. As an end product the system redacts PII elements by substituting the PII type name (e.g., [DATE]) for the text (e.g., 9/11/2001), but for evaluation purpose tagged text is displayed in VTT.

![Figure 1. VTT Window Showing a Hypothetical Annotated Report](image)

The Privacy Rule guidelines published by the Office of Civil Rights of the Health and Human Services (HHS) say that “The importance of documentation for which values in health data correspond to PHI, as well as the systems that manage PHI, for the de-identification process cannot be overstated.” 4
Most studies in the area of automatic de-identification only indicate the set of PII items they redact in the course of de-identification. In their review, Meystre et al. mapped the personal identifiers that were de-identified by the 18 different systems into seven categories: patient names (or both patient and provider names), ages greater than 89, geographical locations, hospitals and healthcare organizations, dates, contact information (phone numbers, pager numbers, fax numbers, and e-mail addresses), and IDs (Social security number, medical record number, driver’s license number, and other identifiers).

The i2b2 challenge uses a list of eight identifier types: patient, doctor, location, hospital, date, ID, phone, and age. Studies based at the Veteran Administration Health System treat a different list of items more in line with their particular needs adding four non-PII categories of clinical eponyms to the i2b2 list. They annotate medical procedure names, medical device names, disease names, and anatomic structures. By supplying the annotations of significant clinical information, they could train their supervised learning system so that it could recognize these entities in text and then evaluate how well the system performs in preserving such clinical information at the end of the de-identification process.

Separating doctor names from patient names is another move in this direction. Although the names of doctors and other medical personnel are not PII, they could not be distinguished with a high level of confidence by automatic de-identification systems and might be redacted inevitably.

In most cases, the details of the annotation scheme are not published. The i2b2 efforts publicly provide their corpus with a data use agreement. This is only possible through an automatic de-identification process followed by an extensive multi-round process of manual validation by human experts. In the resulting corpus the PII elements are substituted with pseudonyms, a surrogate text that looks like the original.

This emphasis on annotation of PII alone overlooks a need for more elaborate annotation, including non-PII items and specific sub-parts of PII items that contribute to evaluation and error analysis and the need to explicitly publish the guidelines used to annotate documents. This paper discusses the annotation scheme we use in the NLM Scrubber project detailing the annotation guidelines and the reasoning behind the decisions that led to this annotation schema.

3. Methods

As mentioned in the previous sections, we have been annotating clinical text since 2008. Our main goal in this effort has always been to develop a set of standards so that we can evaluate the performance of our clinical text de-identification system, NLM Scrubber. Both our annotation and de-identification studies have been influencing and informing each other. While the availability of a standard text has been helping us to test new ideas and monitor the de-identification performance of our system as we modify existing modules, evaluation of the updated de-identification system has also made the shortcomings of our annotation methods explicit. In the following subsections, we describe our annotation methods and explain why we annotate the way we do.

3.1. Annotations on two dimensions

We perceive the annotation space in two orthogonal dimensions. The first dimension denotes personal identifiers. We established a total of 12 personal identifier categories: Address, Personal Name, Personal Name Initials, Organization, Occupation, Telecommunication, Date, Age, Time, Numeric and Alphanumeric Identifiers, Personally Identifying Context, and Role.

The second dimension is personhood, which associates the identifier with an identity. We define 5 personhood categories: Patient, Relative, Employer, Provider, and Other. For example, we would annotate the word “John” in the following two dimensions: It is a personal name and may denote (say) the patient. If the latter is true, we would use the following label PersonalName::Patient. If “John” is the name of the health care provider, we would label it PersonalName::Provider.

We use the personhood category Relative broadly, which includes family members as well as the members of the household of the patient—the Privacy Rule mentions them separately. Given that a family member mentioned in a clinical report is frequently a household member as well, categorizing them separately would be problematic, since we would have to annotate the same word with two distinct personhood labels. Although technical challenges are not insurmountable, it would be conceptually too complex for the annotators to distinguish whether the family member mentioned in the clinical text was also living with the patient in the same house.
Although the Privacy Rule dictates that personal identifiers of the patient’s employer must be de-identified, it does not clarify what constitutes an employer. It could be the owner, president, or the CEO of the company. Could it be the supervisor of the patient? How about their supervisors? In many workplace accident cases, the patient is accompanied to the health care facility by a co-worker. In a re-identification attempt, the co-worker’s identity could be linked to the company and through which, indirectly, to the patient; thus, we use the personhood category Employer to annotate all types of co-workers and supervisors of the patient.

The Provider category denotes every type of healthcare professional who takes part in the health care of the patient. Note that information about the provider was not defined by the Privacy Rule as PII. We use the category Other to denote other personhood identities that are not patients, relatives or providers and there is no apparent method to link that particular person or personal identifier to the patient. For example, we annotate the word Obama in “the patient cited Obama as our president” with the label PersonName::Other. Disclosures of identifiers associated with Provider or Other usually do not pose any significant privacy risk to the patient, since they are not directly linkable to the patient.

How should we annotate girlfriend, partner, and neighbor? We annotate partner as, since it may indicate some kind of formal union and/or household membership, and can be linked to the patient. We use the label Other for friends and other informal relations who may not be linked to the patient directly and as easily as a household member—in the age of social networks, we are not sure how long this assumption would be holding! Although neighbor seems fitting to the label Other at the very first glance, the neighbor information is actually akin to that of the household member, since their residence information could be identifying the address of the patient; thus, we annotate it as Relative.

By reserving the label Other for information that cannot be linked to the patient directly (or indirectly) and by not using it for sensitive information such as information about neighbors, we may prevent significant complications with respect to the evaluation of the de-identification system in case of any unintended disclosure.

In the following subsections, we discuss 12 personal identifier categories, what subcategories, if any, they consist of and how they are related to identifiers mentioned in the HIPAA Privacy Rule. Some entities in these categories may not be personal identifiers. In those cases, we discuss why we chose to introduce and annotate them.

3.2. Address

The Address category comprises a number of entities such as street name, number and types. Table 1 shows which labels we use to annotate such entities. A mention of address may contain a subset of these entities.

<table>
<thead>
<tr>
<th>Label</th>
<th>Entity</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Street</td>
<td>Street name</td>
<td>Pennsylvania Ave</td>
</tr>
<tr>
<td>Location</td>
<td>Street number, apartment, suite or office number; floor or room number inside an office building, hospital or clinic including a bed number, P.O. Box</td>
<td>Station 10-Room 33-A</td>
</tr>
<tr>
<td>Building</td>
<td>Building name</td>
<td>Woodward Building</td>
</tr>
<tr>
<td>City</td>
<td>Village, town or city</td>
<td>Bethesda</td>
</tr>
<tr>
<td>County</td>
<td>County</td>
<td>Montgomery County</td>
</tr>
<tr>
<td>State</td>
<td>State, US district, territory, province or region</td>
<td>D.C. Metro Area, Guam, East Coast, Alberta, Western Pennsylvania</td>
</tr>
<tr>
<td>Country</td>
<td>Country</td>
<td>United States</td>
</tr>
<tr>
<td>Zip</td>
<td>Five or nine digit US ZIP code or foreign postal equivalent</td>
<td>20894-3828, SW1A 2AA</td>
</tr>
</tbody>
</table>

Why do we use eight different address labels, instead of using a single label, to annotate all address tokens? Using a single, common address label sounds quite practical at the first glance, esp. during the annotation process. However, if one needs to assess the performance of a de-identification system that may inadvertently reveal some address information, uniform address labels would be very inadequate for estimating the level of risk to the potential breach
of patient privacy. Note that revealing certain address elements, e.g. a rare street name and number, could pose significantly more risk than revealing more common or widely shared address elements such as an apartment number or name of the city where the patient resides.

HIPAA Privacy Rule makes a distinction between different types of address information. The Privacy Rule states that information about all geographic subdivisions smaller than state, except the first two digits of the zip code, must be de-identified. The third digit of the zip code can be left intact, only if the size of the population in the area of the censored two digits is greater than 20,000 according to the most recent census data. In other words, the Privacy Rule indicates certain address tokens are more informative than others in identifying an individual. If we visualize the address elements on a line ordered from the most granular or specific elements (such as street name and number) towards the most widely shared element (i.e., country), the Privacy Rule puts the threshold between County and State.

If the user intends to fully de-identify patient data, then s/he needs to use the above threshold. However, the Privacy Rule also offers a lower threshold in its Limited Data Set provision, which allows the user to preserve city and town information as long as such information is necessary for the study and the user signs a data use agreement with the provider of the data.

These two thresholds divide the address elements into three parts: If using the Privacy Rule, (A) information more specific than town or city needs to be eliminated under any circumstances; (B) state and country information can be preserved even in a fully de-identified set of data; and (C) information whose specificity lies between these two thresholds that can be preserved only within the boundaries of the Limited Data Set provision. In other words, one needs to use at least three distinct labels to differentiate these three parts in an address. Furthermore, a separate label for ZIP codes is also necessary since ZIP code information crosses these two boundaries.

We could merge State and Country labels into one State/Country label and merge City and County labels into City/County label but we chose not to do so for various reasons. The first reason is practicality—annotating these four types of address elements separately does not impose undue burden onto our annotators. Distinguishing these labels can also be useful under certain situations. For example, in an epidemiologic study in which preserving county information may be necessary and sufficient, de-identifying city information could better protect patient privacy. Unless the user requires state information, NLM Scrubber de-identifies it by default. Although HIPAA does not sanction state information to be de-identified, we choose to do so, since many re-identification algorithms rely on address information and the more unnecessary address information we could de-identify, the more difficult the re-identification would be. Since it is usually difficult to distinguish country of residence from the country of origin, we annotate the residence of a foreign national and the country of the origin of an individual (e.g., “40yo Ethiopian man”) with label Country. We believe a de-identification system ought to preserve ethnicity and country of origin information, since some diseases are more prevalent in certain groups and geographical locations globally.

We distinguish three distinct address elements below the town or city level: Street name, location information such as house or apartment numbers, which further qualify the local address, and building name. Note that inadvertently revealing a street (i.e. house) number without disclosing the street name would not jeopardize privacy of the patient—it would be truly nonspecific. Revealing a street name without the street number however poses a more serious risk, especially if the street name is not a very common one. In that scenario, the privacy risk would be inversely proportional to the size of the household population on all streets with that street name in the country. If on the other hand, the de-identification system inadvertently discloses both street name and number, re-identifying the individual along with age and gender information may not be too difficult. We separated building names from street and location categories, because a building name alone can be more informative than either of them. Note that a building name is at least as informative as the combination of both the street name and the street number. Since it is not customary to name residential units in the US, a building name, as rare as it is, could be quite identifying.

3.3. Personal Names and Personal Name Initials

The Privacy Rule states that names (of the individual or of relatives, employers, or household members of the individual) should be removed. If one has a tabular data where the columns are well defined, it would be easy to distinguish personal names from other identifiers. But when the data in question is text, the seemingly obvious de-identification task can be quite complicated. For example, questions like “would it be okay to leave single letter middle initials intact?” or “would personal name initials constitute a name?” can be difficult to answer.

Clearly, personal name initials like JFK are not as revealing as corresponding full names. In fact, converting names into initials is a widely used practice to protect identities of the minors in reports of the court hearings.1-3 So, we
categorize personal name initials separately from personal names. According to the Office of the Civil Rights, however, personal name initials are considered as personal names and ought to be de-identified.\(^4\) We reserve personal name initials only for the full set of name initials (i.e., when first, middle, and last names are initialized altogether as in JFK) but annotate middle and/or first name initials, as in “James T. Kirk” or “J.K. Rowling,” as parts of the personal names.

Although we annotate suffixes such as Jr. and Sr. as parts of personal names, we do not extend it to professional and academic titles, for some of which we use the label Occupation.

### 3.4 Occupation and Organization

Occupation information is not one of the 18 pieces of PII, sanctioned by HIPAA, to be de-identified. However, especially if it is a rare occupation (e.g., clinical computational linguist, Supreme Court justice), the information may be used to re-identify the patient. Up to date, we have not come up with an easily implementable annotation method to differentiate rare occupation information from the common ones. We have to separate the wheat from the chaff for each piece of occupation information at the evaluation phase of our de-identification studies. Note, however, the personhood dimension that we introduced in this paper for the first time (see Section 3.1) can be helpful when occupation information is associated with Provider or Other, which usually would not pose any privacy risk to the patient.

Most professional titles indicate the occupation of the person. Although we annotate provider occupations (e.g., dermatologist) whenever it is explicitly stated in the text, we have not been annotating their titles (e.g., Dr., M.D., etc.) due to their sheer number of occurrences and the difficulty that it would impose on our annotation team. We are currently studying the feasibility of the issue in a pilot.

We also annotate past occupation information but not the future ones. The former can be linked to the patient but the latter (e.g., “the patient plans to open a car dealership”) is mostly hypothetical. Similarly, we do not annotate hobbies as occupations since they would rarely be unique and linkable to the patient. In such rare scenarios, however, we have other methods to employ (see Section 3.7).

Occupation (e.g. a cook) does not specify the employer like where the person works (e.g., “… at Acme Restaurant”), but sometimes, they are very closely linked together. For example, “he is an Army Master Sergeant,” where we annotate Army with label Organization::Employer and Master Sergeant with Occupation::Patient or Occupation::Relative, depending on whom “he” denotes. If the title were Admiral, for which we would use label Occupation::Patient, it would also implicitly reveal the employer’s organization, Navy.

We reserve the personhood label Employer only for the patient’s employer and do not extend it to the employer of the relative, since there is no apparent direct link from the employer to the patient. In the example, “The patient’s mother is a math teacher at Takoma Park Middle School,” math teacher is Occupation::Relative and Takoma Park Middle School is Organization::Relative. Between the school and the patient, there is two degrees of separation, which is implied by the label Organization::Relative—the linkage for re-identification is possible but the link is weaker than the link between the patient and their employer.

Although we do not annotate hobbies, we do annotate organizations that individuals can be associated with (e.g., “the patient is a member of the Rotary Club” or “…presented his findings during the AMIA Symposium last year”).

### 3.5 Age, Date and Time

Similar to category Address, Age and Date are categories, each of which comprises multiple labels. By mandating that ages over 89 be de-identified, HIPAA separates age into two categories: (1) ages 90 and above are considered PII, which we annotate with label AgePII, and (2) ages that are below 90, which HIPAA considers as non-PII. We split the second group into two additional separate groups: (2A) ages that are mentioned as whole numbers, which we annotate with label AgeNPII, and (2B) ages that are mentioned as fractions of whole years (e.g., “Patient is a 4 and 11/12 month old boy”), which we annotate with label AgeFraction.

Without an anchor to a fixed date AgeFraction is not very useful to re-identify the patient; thus, it should be considered as non-PII. However, it is possible that a de-identification system might miss a mention of the report date, which, along with the age information in fractions (e.g., “he will be 11 months old in two days”), one may be able to identify
the birth date of the patient. In other words, label AgeFraction could pose privacy risk only in conjunction with an inadvertently revealed full-date within the text.

If the patient’s current age is 90 or older and the narrative report provides indirect reference to the patient’s age such that re-identification can be done through a simple arithmetic (e.g., “Twenty years ago, at the age of 75, he had an ischemic attack”), we would annotate the earlier age references (i.e., 75 in the example above) as AgePII as well.

We do not annotate other “age” types such as gestational age, bone age (unless identical to the chronological age), school grade level (10th grade) or age periods such as teenage, middle-aged, etc., since they are not as identifying as chronological age found in formal records.

The category Date comprises six labels: Year (e.g., 2001), Month (e.g., September), Day (e.g., 11th), DayOfWeek (e.g., Tuesday but not Tuesdays), SpecialDay (e.g., 9/11, Hurricane Sandy, Katrina, Cinco de Mayo, New Year), and Period (e.g., flu season, Monsoon, Ramadan, winter, second trimester).

We annotate not only those special days that are fixed in history such as Pearl Harbor, 2008 Market Crash but also those special days that occur every year such as New Year, whose exact dates can be construed when combined with year information, which taken alone is not PII under HIPAA. We also label personal special days such as birthday or Bar Mitzvah, not only due to potential privacy concerns as they may be available from external sources, but also due to their potential importance in reference to other events in the narrative text.

We use the label Period to annotate any time period longer than a day of which begin and end dates are not explicitly stated. We use this label to annotate periods in the patient’s medical history such as pregnancy, puberty, hospitalization period, and menstruation as well as calendar periods such as early 2001 or in the 90s. Most age references in the medical history are periods. For example, “when the patient was 5 years old …” or “spoke at 5–5 ½ years old”. Note that age references in these examples do not denote the patient’s current age but if such age references in the past reveal that the patient’s current age is 90 or older, we would have to use label AgePII instead.

If a period of two days or longer is described in terms of an interval or a range with explicit begin and end date identifiers (e.g., 1995–97, between next Tuesday and Friday), we separately annotate begin and end points with the appropriate date label. If it is an age range, we label each age separately. In the example, “had hearing loss from 85-97 years old”, we annotate 85 with AgeNPII and 97 with AgePII.

Recall that we define the Period as a subcategory of date; therefore, we use it only if the period can be stated relative to a date. In example, “when the patient was 5 years old”, we perceive a period of one year, starting 5 years after the birth date. If the period is stated using terms like last year, last month, last week, and last weekend, the period is defined relative to the date of the report. We do not annotate (hence do not use the label Period) cyclical temporal references such as daily, Tuesdays or every Tuesday or other temporal references described in sequence of events without any apparent date to anchor (e.g., “completed 2 weeks of antibiotics”).

We annotate last Christmas or Christmas last year as SpecialDay since the terms last and last year further qualify the special day, but when the year is explicitly stated as in Cinco de Mayo 2000, we annotate Cinco de Mayo as SpecialDay and annotate 2000 as Year, because in this example, the date term refers to a full date May 5, 2000.

We do annotate time of the day using the label Time, but we also believe that it is too general to link to the patient for re-identification. Since we do not classify it under the Date category, we do not annotate time periods within a day as Period (e.g., noon-4:30pm); instead, we use label Time to annotate noon and 4:30pm, separately.

3.6. Telecommunication and Alphanumeric Identifiers

Telecommunication identifiers are the most straightforward and the least ambiguous identifiers since they are well defined engineering objects. Of the 18 personal identifiers defined by the HIPAA Privacy Rule, five of them are telecommunication identifiers to be de-identified: telephone numbers, fax numbers, electronic email addresses, web universal resource locators (URLs), and Internet protocol (IP) address numbers. As new telecommunication modes and media emerge, new telecommunication identifiers (e.g., Twitter usernames such as @BarackObama) appear, but they too are covered by the last (18th) catch-all identifier of the Privacy Rule: “any other unique identifying number, characteristics, or code” must be de-identified.

Numeric and Alphanumeric Identifiers consist of four labels: MedicalRecordNo, ProtocolNo, HealthRecordID, and AlphanumericID. The first two are very specific identifiers denoting medical record and protocol numbers,
respectively. Medical record number is one of the 18 HIPAA identifiers. Since protocol numbers are very important entities for clinical researchers, who are the intended users of NLM Scrubber, we annotated them separately. We use the label HealthRecordID for all other alphanumeric identifiers issued by health care and insurance providers uniquely to the patient; e.g., hospital account number, health plan beneficiary number and lab specimen number. MedicalRecordNo, ProtocolNo and HealthRecordID are almost always associated with the patient—only in a handful of cases did we observe mentions of such identifiers for the relatives.

We use the label AlphanumericID for all other numeric and alphanumeric identifiers that are not issued by the provider, including those five identifiers defined by the Privacy Rule: social security number, account numbers, certificate/license numbers, vehicle identifiers, and device identifiers. Note for hospital account numbers we use the label HealthRecordID.

Sometimes, names of some lab materials and experimental drugs may contain some numbers (e.g., drug 123-ABC or instrument QRS-40). We do not annotate such health information, as they are neither unique to the patient nor personal identifiers.

3.7. Personally Identifying Context

So far, we discussed how we annotate entities that were mentioned in the HIPAA Privacy Rule along with a few other closely related entities, some of which can be PII in certain contexts. We are aware of the fact that due to the intricacies of natural languages, it is possible to specify a context in which the person could be identified indirectly such that no labels we discussed so far would be appropriate to use. In those cases, we label the tokens with PIC, denoting Personally Identifying Context.

In the hypothetical example, “received his injuries while he was reporting from Tahrir Square”, we would annotate reporting with label Occupation::Patient and Tahrir Square with PIC::Patient, since the latter would provide context so specific that along with the occupation information would probably identify the person directly.

In the example “the patient was deployed to Iraq”, the reader may presume that the patient is in the military, but the deployment to Iraq is not an occupation—equipment could be deployed to Iraq as well as other types of personnel such as reporters could be deployed to a war zone. However if the example provides an occupational context that is so specific that it might tighten the circle of potential candidates, we would label those tokens as PIC. But in this example, even if we presume that the context alludes that the subject is a military person, the circle of military personnel remains too broad to label the phrase as PIC.

3.8. Role

In order to associate a personal identifier with a person, automatic de-identification system needs to recognize a reference to that person. We define such a reference as Role, which can denote the patient, mother, father, daughter, supervisor, physician, boyfriend, and others. We annotate those roles in order to evaluate and monitor our system’s performance. Although they too are roles, we do not annotate pronouns such as he, she, him, hers, their, themselves etc. We use the label Role only if no other label is suitable for that annotation. For example, if the provider’s occupation is more specific than the role of physician or nurse, such as cardiologist or physical therapist, then we annotate it as Occupation. If the reference specifies a personally identifying context, instead of using the label Role, we would annotate it as PIC.

The role information is quite important in the context of the deceased patient records as well, because even though health records of the deceased patient may not constitute protected health information, health information of their living relatives does. Fortunately, such information is quite rare. Recognizing such roles in the narrative reports of the deceased helps prevent such privacy breaches.

4. Results

Our annotation label set and methods of annotating text elements that we described in this paper are the results of the seven years long evolution of annotation, de-identification, and evaluation. By defining the annotation labels on two dimensions and associating identifiers with personhood, Patient, Relative, Employer, Provider, and Other, we can easily stratify the importance of text elements in terms of high, medium, low, and no privacy risks.
We divided some identifier categories such as Address into subcategories, each with a distinct label. Even though
some information (e.g., house or street numbers labeled with Location) seem more granular or specific than others
(e.g., town labeled with City), inadvertently revealing them would pose little or no privacy risk; however such
identifiers (e.g., house number and street name) become very significant only if they are revealed in combination with
certain other elements of the same category (e.g., house number and street name together). The same is true for the
subcategories of Date; i.e., day, month, or year information alone has no significance until they are revealed together.

The newly introduced special subcategories and associated labels such as Period, SpecialDay, and AgeFraction enrich
our label set and provide clarity and direction to our annotators when faced with non-standard and borderline cases.
For example, “at age 3, the patient started…” may seem to contain a piece of information about the patient’s age at
the very first glance, but “age 3” actually outlines a period in the medical history of the patient and does not identify
how old the patient currently is. In short, these new labels yield a corpus with more accurate annotations.

Personally Identifying Context labeled with PIC is a very important new category since we no longer need to say
“There could be some information identifying the patient indirectly without using any explicit PII elements in this
report.” Now, if we encounter such information, we have the tool to annotate it.

5. Discussion

In this paper, we introduced a new annotation schema that extends the identifier elements of the HIPAA Privacy Rule.
In this schema, we annotate text elements on two dimensions: identifier type and personhood denoted by the identifier.
The personhood can take one of the following type values: Patient, Relative, Employer, Provider and Other. We
extended identifier types both in terms of scope and granularity.

Our annotation label set is based first and foremost on the PII elements defined by the HIPAA Privacy Rule. However,
being aware of other annotation efforts, we tried to design a broad spectrum of annotation labels so that we can
establish a common ground for our community. Standardization of annotation schemas is a very important goal that
we all should strive for; otherwise, an effective evaluation and comparison of our study results would be too difficult.
We believe this is the first step towards that ambitious goal.

The concepts and annotation methods defined and described in this paper could be best understood if studied along
with a number of good examples. We are currently working on finalizing our annotation guidelines containing a rich
set of examples most of which are extracted from actual reports. The guidelines will be publicly available by the time

Acknowledgements

We are grateful to Brett South, Guy Divita and their colleagues for sharing with us the annotation guidelines used in
their research at the University of Utah and the VA Salt Lake City Health Care System.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library
of Medicine.

Competing Interests

The first author receives royalties from University of Pittsburgh for his contribution to a de-identification project.
NLM’s Ethics Office reviewed and approved his appointment.

References


Clinical Predictive Modeling Development and Deployment through FHIR
Web Services

Mohammed Khalilia, Myung Choi, Amelia Henderson, Sneha Iyengar, Mark Braunstein, Jimeng Sun
Georgia Institute of Technology, Atlanta, Georgia

Abstract
Clinical predictive modeling involves two challenging tasks: model development and model deployment. In this paper we demonstrate a software architecture for developing and deploying clinical predictive models using web services via the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) standard. The services enable model development using electronic health records (EHRs) stored in OMOP CDM databases and model deployment for scoring individual patients through FHIR resources. The MIMIC2 ICU dataset and a synthetic outpatient dataset were transformed into OMOP CDM databases for predictive model development. The resulting predictive models are deployed as FHIR resources, which receive requests of patient information, perform prediction against the deployed predictive model and respond with prediction scores. To assess the practicality of this approach we evaluated the response and prediction time of the FHIR modeling web services. We found the system to be reasonably fast with one second total response time per patient prediction.

Introduction
Clinical predictive modeling research has increased because of the increasing adoption of electronic health records\(^1\)\(^-\)\(^3\). Nevertheless, the dissemination and translation of predictive modeling research findings into healthcare delivery is often challenging. Reasons for this include political, social, economic and organizational factors\(^4\)\(^-\)\(^6\). Other barriers include the lack of computer programming skills by the target end users (i.e. physicians) and difficulty of integration with the highly fragmented existing health informatics infrastructure\(^7\). Additionally, in many cases the evaluation of the feasibility of predictive modeling marks the end of the project with no attempt to deploy those models into real practice\(^8\). To achieve real impact, researchers should be concerned about the deployment and dissemination of their algorithms and tools into day-to-day decision support and some researchers have developed approaches to doing this. For example, Soto \textit{et al.} developed EPOCH and ePRISM\(^7\), a unified web portal and associated services for deploying clinical risk models and decision support tools. ePRISM is a general regression model framework for prediction and encompasses various prognostic models\(^9\). However, ePRISM does not provide an interface allowing for integration with existing EHR data. It requires users to input model parameters, which can be time consuming and a particular challenge for researchers unfamiliar with the nuances of clinical terminology and the underlying algorithms. A suite of decision support web services for chronic obstructive pulmonary disease detection and diagnosis was developed by Velickovski \textit{et al.}\(^10\), where the integration into providers' workflow is supported through the use of a service-oriented architecture. However, despite these few efforts and many calls for researchers to be more involved in the practical dissemination of their systems, little has been done and much less has been accomplished to utilize predictive modeling algorithms at the point-of-care.

An important missing aspect that retards bringing research into practice is the lack of simple, yet powerful standards that could facilitate integration with the existing healthcare infrastructure. Currently, one major impediment to the use of existing standards is their complexity\(^11\). The emerging Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) standard provides a simplified data model represented as some 100-150 JSON or XML objects (the FHIR Resources). Each resource consists of a number of logically related data elements that will be 80% defined through the HL7 specification and 20% through customized extensions\(^12\). Additionally, FHIR supports other web standards such as XML, HTTP and OAuth. Furthermore, since FHIR supports a RESTful architecture for information and message exchange it becomes suitable for use in a variety of settings such as mobile applications and cloud computing. Recently, all the four major health enterprise software vendors (Cerner, Epic, McKesson and MEDITECH) along with major providers including Intermountain Healthcare, Mayo Clinic and Partners Healthcare have joined the Argonaut Project to further extend FHIR to encompass clinical documents constructed from FHIR resources\(^13\). Epic, the largest enterprise healthcare software vendor, has a publicly available FHIR server for testing that supports a subset of FHIR resource types including Patient, Adverse Reaction, Medication Prescription, Condition and Observation\(^14\) and support of these resources is reportedly included in their June 30, 2015 release of Version 15 of their software. SMART on FHIR has been developed by Harvard Boston Children’s Hospital as a
A universal app platform to seamlessly integrate medical applications into diverse EHR systems at the point-of-care. Cerner and four other EHR systems demonstrated their ability to run the same third party developed FHIR app at HIMSS 2014. Of particular importance to our work is the demonstrated ability to provide SMART on FHIR app services within the context and workflow of Cerner’s PowerChart EHR.

As a result of these efforts, FHIR can both facilitate integration with existing EHRs and form a common communication protocol using RESTful web services between healthcare organizations. This provides a clear path for the widespread dissemination and deployment of research findings such as predictive modeling in clinical practice. However, despite its popularity, FHIR currently is not suitable to directly support predictive model development where a large volume of EHR data needs to be processed in order to train an accurate model. To streamline predictive model development it is important to adopt a common data model (CDM) for storing EHR data. The Observational Medical Outcomes Partnership (OMOP) was developed to transform data in disparate databases into a common format and to map EHR data into a standardized vocabulary. The OMOP CDM has been used in various settings including drug safety and surveillance, stroke prediction, and prediction of adverse drug events. We utilize a database in the OMOP CDM to support predictive model development. The resulting predictive model is then deployed as FHIR resources for scoring individual patients.

Based on these considerations, in this paper we propose to develop and deploy:

- A predictive modeling development platform using the OMOP CDM for data storage and standardization
- A suite of predictive modeling algorithms operating against data stored in the OMOP CDM
- FHIR web services, that use the resulting trained predictive models to perform prediction on new patients
- A pilot test using MIMIC2 ICU and ExactData chronic disease outpatient datasets for mortality prediction.

Methods

Overview and system architecture

Figure 1 shows how our architecture supports providing predictive modeling services to clinicians via their EHR. On the server side, the model development platform trains and compares multiple predictive models using EHR data stored in the OMOP CDM. After training, the best predictive models are deployed to a dedicated FHIR server as executable predictive models. On the client side, users can use existing systems such as desktop or mobile applications within their current workflows to query the predictive model specified in the FHIR resource. Such integrations are done by using FHIR web services. Client applications use FHIR resource to package patient health information and transport it using the FHIR RESTful Application Programming Interface (API). Once the FHIR server receives the information, it passes it on to the deployed predictive model for the risk assessment. The returned result from the predictive model will be sent to the client and also stored into a resource database that can be accessed by the client to read or search the Risk Assessment resources for later use.

![Figure 1. System Architecture](image-url)
The common data model

Developing a reliable and reusable predictive model, requires a common data model into which diverse EHR data sets are transformed and stored. For the proposed system we used the OMOP CDM designed to facilitate research using some important design principles. First, data in the OMOP CDM is organized in a way that is optimal for data analysis and predictive modeling, rather than for the operational needs of healthcare providers and other administrative and financial personnel. Second, OMOP provides a data standard using existing vocabularies such as the Systematized Nomenclature of Medicine (SNOMED), RxNORM, the National Drug Code (NDC) and the Logical Observation Identifiers Names and Codes (LOINC). As a result predictive models built using data in an OMOP CDM identify standardized features, assuming the mapping to OMOP is reliable. The OMOP CDM is also technology neutral and can be implemented in any relational database such as Oracle, PostgreSQL, MySQL or MS SQL Server. Third, our system can directly benefit the existing OMOP CDM community to foster collaborations.

The OMOP CDM consists of 37 data tables divided into standardized clinical data, vocabulary, health system data and health economics. We only focused on a few of the CDM clinical tables including: person, condition occurrence, observations and drug exposure. As we enhance our Extract, Transform, Load (ETL) process and the predictive model, we can incorporate additional data sources as needed. Figure 2 shows a high level overview of the ETL process in which multiple raw EHR data are mapped to their corresponding CDM instances. In the transformation process, EHR source values such as lab names and results, diagnoses codes and medication names are mapped to OMOP concept identifiers. The standardized data can then be accessed to train the predictive models.

Predictive model development

The CDM provides the foundation for predictive modeling. As various datasets are transformed and loaded into the OMOP CDM predictive modeling training can be simplified because the OMOP CDM instances all have the same structure. For instance, one can equally easily train a model for predicting mortality, future diseases or readmission using different datasets. Figure 3 provides an overview of predictive model training.
The predictive models are trained offline and can be re-trained as additional records are added to the database. Training consists of three modules: 1) **Cohort Construction**: this is the first step in the training phase. At this stage the user specifies the OMOP CDM instance, prediction target (i.e. mortality) and the cohort definition. Based on the specified configuration the module will generate the patient cohort. 2) **Feature Construction**: at this stage the user specifies which data sources (e.g. drugs, condition occurrence and observations) to include when constructing the features. Additional configurations can also be provided for each data source. The user can include the observation window (e.g. the prior year, to utilize only patient data recorded in the past 365 days). Other data source configurations include the condition type concept identifier to specify which types of conditions to include (i.e. primary, secondary), observation type concept identifier and drug type concept identifier. The final configuration is the feature value aggregation function. For example, lab result values can be computed using one of five aggregation functions: sum, max, mean, min, and most recent value. 3) **Model Training**: This module takes the feature vectors constructed for the cohort and trains multiple models using algorithms such as Random Forest, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). The parameters for each of these three algorithms are tuned using cross validation in order to select the best performing model. The best model will be deployed as a FHIR resource. In the next section, we will describe how to deploy such a predictive model for scoring future patient in real time.

**FHIR web services for model deployment**

Our approach is focused on API based predictive modeling services that can be easily implemented in thin client applications, especially in the mobile environment. FHIR defines resources represented as JSON or XML objects that can contain health concepts along with reference and searchable parameters. FHIR further defines RESTful API URL patterns for create, read, update, and delete (CRUD) operations. In this paper, we propose to use the RiskAssessment resource defined in the FHIR Draft Standard for Trial Use 2 (DSTU2) for our predictive analysis. Readers should note that this particular FHIR resource is still in the draft stage. The current approved DSTU1 version of FHIR does not have a RiskAssessment resource. However, our development version of FHIR is currently being balloted on by HL7 members and should be approved soon. Detailed information about the FHIR development version for RiskAssessment can be obtained from Ref. 20.

A prediction request to the FHIR server starts by forming the CREATE operation, which is used to request for scoring specific patients in real-time using the deployed predictive model. This creates RiskAssessment resources at the server. Client applications then receive a status response with a resource identifier that refers to the newly created resource. Clients can use this resource identifier to read or search the resource database via a FHIR RESTful API. For this paper, we used a SEARCH operation as we need to retrieve more than one result. Groups of FHIR resources are called a bundle and there is a specified format for that. As all performed analyses are stored in the resource database, additional query types can be implemented in the future. The operation process is depicted in Figure 4.

![Figure 4. Operation Process](image)

During the process, clients and server need to put appropriate information into elements available in the RiskAssessment resource. However, most of the data elements in the RiskAssessment resource are optional for predictive modeling which gives us flexibility in choosing the model output. For the CREATE operation, the RiskAssessment resource is constructed with subject, basis, and method elements as shown in Figure 5(a).
Subject is used to define a group identifier, which we also refer to as the resource identifier. Basis can contain information used in assessment, which are the patient identifiers. All patients included in basis are bound to the group identifier specified in the subject element. Therefore, the results of assessment for patients will be grouped in the resource database by the group identifier at the FHIR server. Subject or group identifier provides a simple mechanism for users to specify a group of patients, whose prediction score can be retrieved from the resource database.

Predictive model scoring creates feature vectors (one for each patient), which are derived from the patient’s health information. The FHIR Patient resource can be referenced by either the patient identifier or by a collection of resources such as MedicationPrescriptions, Conditions, Observations, etc. In our initial implementation we used patient identifier. Using the patient identifier, the predictive model constructs feature vectors by pulling the appropriate information from various resources that contain each patient’s clinical data occurring within the specified observation window. In our implementation the clinical data is stored in the OMOP database. This all happens during the predictive analysis period in Figure 4. In the future, we will construct features directly from EHR databases through querying other FHIR resources in real-time.

Results from the predictive analysis are sent back to the FHIR server in JSON format and the FHIR server stores the information in a resource database. If the CREATE operation is successful, the server will send a 201-status created message. In case of any errors, an appropriate error status message will be sent back to the client with the OperationOutcome resource, if required.

Clients can query the prediction results using the resource identifier returned in the CREATE response (i.e. group identifier under subject in this case). This resource identifier can be used to query the stored prediction results using the SEARCH operation. Recall that the resource identifier is bounded to a list of patient identifiers. Once the SEARCH operation request is received with the resource identifier, the FHIR server constructs a response FHIR resource for each patient. In the RiskAssessment resource, the prediction element contains the risk score for a patient, as shown in Figure 5(b). When the SEARCH operation is completed, the RiskAssessment resources are packaged into a bundle and sent to the client. For example, the resource contains an identifier that represents the patient for whom the predictive analysis is performed. In our experiments, mortality prediction was performed and risk scores are returned. The mortality scores are populated in the probabilityDecimal sub element of the prediction element, as shown in Figure 5(b).

Experiments

We tested our implementation using two datasets: Multiparameter Intelligent Monitoring in Intensive Care (MIMIC2) and a dataset licensed from ExactData. MIMIC2 contains comprehensive clinical data for thousands of ICU patients collected between 2001 and 2008. ExactData is a custom, large realistic synthetic EHR dataset for
chronic disease patients. Using ExactData eliminates the cost, risk and legality of real EHR data. Table 1 presents key statistics about the two datasets.

Table 1. Key ExactData and MIMIC2 Dataset Statistics

<table>
<thead>
<tr>
<th></th>
<th>ExactData</th>
<th>MIMIC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>10,460</td>
<td>32,074</td>
</tr>
<tr>
<td>Number of condition occurrences</td>
<td>313,921</td>
<td>314,648</td>
</tr>
<tr>
<td>Number of drug exposures</td>
<td>82,258</td>
<td>1,736,898</td>
</tr>
<tr>
<td>Number of observations</td>
<td>772,189</td>
<td>19,272,453</td>
</tr>
<tr>
<td>Number of deceased patients</td>
<td>53</td>
<td>8,265</td>
</tr>
</tbody>
</table>

Two ETL processes were implemented to move the raw MIMIC2 and ExactData datasets into two OMOP CDM instances. From each instance a cohort was generated for mortality prediction. The cohort for ExactData is limited to the 53 patients with death records matched with 53 control patients to keep it balanced. The MIMIC2 cohort is somewhat larger with 500 case patients and 500 control patients. To generate the control group we performed a one-to-one matching with the case group based on age, gender and race. As illustrated in Figure 3, the feature vectors for each cohort are generated using these MIMIC2 and ExactData cohorts. The features include condition occurrence, observation and drug exposures. The observation window was set to 365 days for both cohorts and the feature values for observations were aggregated by taking the mean of the values.

For each cohort three mortality predictive models were trained offline using Random Forest, SVM and KNN. After performing cross validation and parameter tuning for each algorithm, the model with the highest Area Under the Receiver Operating Characteristic curve (AUC) was deployed. This results in three final models, one for each algorithm, allowing clients to specify which algorithm to use to predict new patients. The predictive models were trained using Python scikit-learn machine learning package. The final model training runtime, AUC, accuracy and F1 score are reported in Table 2.

Table 2. Model evaluation using ExactData and MIMIC2 cohorts

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Runtime (seconds)</th>
<th>AUC</th>
<th>Accuracy</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>2.5</td>
<td>0.87</td>
<td>0.75</td>
<td>0.76</td>
</tr>
<tr>
<td>SVM</td>
<td>2.5</td>
<td>0.90</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>Random Forest</td>
<td>2.52</td>
<td>0.95</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>MIMIC2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>392</td>
<td>0.71</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td>SVM</td>
<td>387</td>
<td>0.80</td>
<td>0.63</td>
<td>0.77</td>
</tr>
<tr>
<td>Random Forest</td>
<td>404</td>
<td>0.78</td>
<td>0.74</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Our entire software platform was evaluated and tested on a small Linux server with 130GB hard drive, 16GB memory, and two Intel Xeon 2.4GHz processors with two cores each. FHIR web services were built on the Tomcat 8 application server using the HL7 API (HAPI) RESTful server library. The FHIR web service expects a request from the client to perform patient scoring (in our case mortality prediction) with patient data or patient identifier contained in the request body. We evaluated the performance of the web service in two ways. The first evaluation method measures the response time from the client’s request to the web service until a response is received back from the server. This is an important measure since a very slow response time makes the system unusable for practical applications including at the point-of-care where busy clinicians demand a prompt response. A single API request to score a patient is a composite of two different requests: 1) CREATE request made to the FHIR server that sends the data in the JSON format. The data comprises of resource type, a set of patient identifiers, classification algorithm to be employed (e.g. KNN, SVM or Random Forest) and the OMOP CDM instance to be used (e.g. MIMIC2 or ExactData). The response time of this request increases with the number of patients (as can be observed from Figures 6 and 7). SEARCH request is made by passing the patient group identifier as a parameter to obtain the prediction for the patients. Similar to the CREATE request, a larger number of patients increases the response time of the SEARCH request.
Figure 6 shows the FHIR web service response time using KNN and MIMIC2 datasets as we vary the number of patients to score in the request. For a single patient, a typical point-of-care scenario, the response time is around one second. The CREATE request consumes most of the time and the response time increases with the number of patients. This could become an issue in an application such as population health to identify those chronic disease patients most likely to be readmitted or screening an entire ICU to identify patients most in need of immediate medical attention. The SEARCH request has a relatively constant response time and does not change significantly as the number of patients increases.

The actual scoring or prediction of the patient is performed in the CREATE request. There are two major tasks that take place in this CREATE request. First is the feature construction that is performed for each patient included in the request. This requires querying the OMOP CDM database, extracting and aggregating feature values, which can be an expensive operation when scoring a large number of patients. The second task is the actual prediction operation which takes the least amount of time. Figure 7 shows the response time (after subtracting the feature construction and prediction time), feature construction time, and prediction time as we increase the number of patients. As the number of patients increases the prediction time does not increase much compared to the response time or the feature construction time. For instance, a request for scoring 200 patients takes about 23 seconds, out of which 8 seconds were spent constructing features and only 16 milliseconds were spent on the actual predictions.
The second web service performance evaluation measures the amount of load that the service can handle within some defined time interval. As this service handles multiple clients it is important to guarantee that it is available and able to respond in a timely manner to requests as the number of clients increases. However, keep in mind that our evaluation is only a proof of concept done on a moderate server. Figure 8 shows the average response time when 1000 clients send CREATE requests (one patient in each request) to the FHIR server within a minute. Figure 8 is generated by the Simple Cloud-based Load Testing tool and has three axes: response time on y-axis, number of clients on secondary y-axis and time on x-axis, where 00:05 indicates the fifth second from the start of the test. The green curve (upper curve) shows the number of simultaneous clients sending the requests at some point in time. For instance, on the fifth second, about 50 clients were sending simultaneous requests. The blue line shows the average response time for those clients (on the fifth second, the average response time for 50 clients was about 1800 milliseconds). Overall, the average response time for the 1000 clients is 1506 milliseconds (about 1.5 seconds). A 1.5 second response time is acceptable when the task is performed at the point-of-care.

![Figure 8. Average response time as 1,000 clients send CREATE request within one minute](image)

**Conclusion and Discussion**

In this paper we presented a real-time predictive modeling development and deployment platform using the evolving HL7 FHIR standard and demonstrated the system using MIMIC2 ICU and ExactData chronic disease datasets for mortality prediction. The system consists of three core components: 1) The OMOP CDM which is more tailored for predictive modeling than healthcare operational needs and stores EHR data using standardized vocabularies. 2) Predictive model development, which consists of cohort construction, feature extraction, model training and model evaluation. This training phase is streamlined, meaning that it will work for any type of data stored in an OMOP CDM. 3) FHIR web services for predictive model deployment that we use to deliver the predictive modeling service to potential clients. The web service takes a prediction request containing patient identifiers whose features can be extracted from the OMOP database. The prediction or the scoring is not pre-computed but is performed in real-time. Our future work on the FHIR web services will enhance the feature extraction by accessing EHR data over FHIR Search operations. In this case, our predictive platform will become a client to the FHIR-enabled EHR. Recent FHIR ballot for DSTU2 includes a Subscription resource. This resource, if included in the standard, can be utilized in our platform to subscribe the patient’s feature related data from their EHR. The Subscription resource uses Search string for its criteria, thus our future work will comply with this new resource with only a few modifications. Even if this resource couldn’t be included, our future platform will use a pull mechanism to retrieve EHR data for feature extraction.

A total of six predictive models were trained on MIMIC2 ICU and ExactData chronic disease datasets from which we generated cohorts based on patients with a death event and a matched set of control patients. The FHIR web service routes the incoming requests to the desired predictive model, which is usually a client specified parameter. For a practical, real-time web service it is important to achieve a fast response time. The observed response time for scoring one patient was around one second, of which the actual prediction took only few milliseconds. The total response time increases with the number of patients in a single request, but the actual prediction time remains very
small, reaching 16 milliseconds when scoring 200 patients. However, scoring many patients in one request is not always desired or needed. For instance, in many cases providers are only interested in querying one patient at a time from the point-of-care. In such a direct care delivery scenario only a single request containing one patient’s information would be sent to the server with an expected response time of one second. Additionally, this response time can be easily reduced by expanding the server hardware and optimizing the implementation.

We present a prototype with much work and many needed improvements yet to be done. First, additional predictive algorithms should be added. We demonstrated Random Forest, SVM and KNN, but others such as Logistic Regression and Decision Trees could be added. Second, the system should allow for updating the predictive model as new patients get scored. This requires using the patient EHR data passed into the web service for retraining the predictive model. For this to take place, the web service should allow the provider to give feedback on the patient scoring, which can then be used to improve future predictions. Third, the system needs to be scalable and should handle larger datasets to train the predictive models. This can be done by utilizing big data technologies such as Hadoop or Apache Spark. Fourth, the response time needs to be improved especially for requests that contain large numbers of patients, which can be done by allocating more resources and improving the FHIR web service software. Additionally, the response time can be decreased if the client includes patient EHR data in the request, thus avoiding expensive database querying. Fifth, the web service must have privacy and security protocols implemented such as OAuth2, which has been already implemented in the SMART on FHIR app platform.

The proposed approach in this paper assumes that the client that is deploying our predictive model stores the patient data in an OMOP CDM model and the CREATE request contains the identifier for the patient for which the analysis is to be performed. Any ETL process for moving data from EHR to OMOP model is outside the scope of this paper. However, FHIR resources such as MedicationPrescriptions, Conditions, Observations can be included in the body of the CREATE request instead of the patient identifier. This allows for performing analysis for patients not included in the database. Additionally, two processes can be performed on the FHIR resources, if passed in the CREATE request: 1) perform prediction using the patient data (resources) in the request and 2) subscribe the patient’s features related data from EHR to the predictive platform.

In conclusion, we have demonstrated the ease of developing and deploying a real-time predictive modeling web service that uses open source technologies and standards such as OMOP and FHIR. With this web service we can more easily and cost effectively bring research into clinical practice by allowing clients to tap into the web service using their existing EHRs or mobile applications.

Acknowledgment

This work was supported by the National Science Foundation, award #1418511, Children's Healthcare of Atlanta, CDC i-SMILE project, Google Faculty Award, AWS Research Award, and Microsoft Azure Research Award.

References


An Ensemble Method for Spelling Correction in Consumer Health Questions

Halil Kilicoglu, PhD, Marcelo Fiszman, MD, PhD, Kirk Roberts, PhD,
Dina Demner-Fushman, MD, PhD
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine
Bethesda, MD

Abstract

Orthographic and grammatical errors are a common feature of informal texts written by lay people. Health-related questions asked by consumers are a case in point. Automatic interpretation of consumer health questions is hampered by such errors. In this paper, we propose a method that combines techniques based on edit distance and frequency counts with a contextual similarity-based method for detecting and correcting orthographic errors, including misspellings, word breaks, and punctuation errors. We evaluate our method on a set of spell-corrected questions extracted from the NLM collection of consumer health questions. Our method achieves a $F_1$ score of 0.61, compared to an informed baseline of 0.29, achieved using ESpell, a spelling correction system developed for biomedical queries. Our results show that orthographic similarity is most relevant in spelling error correction in consumer health questions and that frequency and contextual information are complementary to orthographic features.

1 Introduction

Orthographic errors are pervasive in informal writing. The questions that consumers ask about their or someone else’s health often contain many misspellings\textsuperscript{[1]}. Misspellings may not pose a significant cognitive burden for a human reader, but they can severely limit the effectiveness of an automated system. At the National Library of Medicine (NLM), we have been building a system to assist customer service staff in answering health questions received from consumers. In 2014, we received more than 40K questions, approximately 15% of which sought health-related information. Our system currently uses a combination of rule-based and supervised machine learning techniques for question understanding. More specifically, we extract focus (generally a disease)\textsuperscript{[2]} and question type (e.g., treatment, prognosis)\textsuperscript{[3]} from the question and construct a semantic frame\textsuperscript{[4]}, which is then converted to a search engine query. These techniques assume well-written questions; thus, orthographic errors significantly hinder their performance. For instance, consider the following question:

(1) My mom is 82 years old suffering from anixity and depression for the last 10 years was dianosed early on set deminita 3 years ago. Do yall have a office in Greensboro NC? Can you recommend someone. she has seretona syndrome and nothing helps her.

Four disorders are mentioned in the question, three of which are misspelled (anixity for anxiety, deminita for dementia, and seretona syndrome for serotonin syndrome). Another, perhaps less central but potentially important, misspelling is dianosed for diagnosed. On the other hand, the misspelling of the colloquial yall (y’all) and nothing (nothing) may be less significant for interpretation of this question. In contrast to such non-word spelling errors, the spelling of onset as on set constitutes a real word spelling error as well as a word break error, since both on and set are valid words. Real word spelling errors can be even more problematic, as in the question “What can I do to lesson the severity of the adema?”, where the underlined words, both crucial for understanding the question, are misspelled. Without detecting and correcting these...
types of errors, there is little hope of extracting the information needed from this question to answer it automatically.

Punctuation errors can also have an impact on question understanding. Omitting sentence-ending periods or space after punctuation, for example, are likely to cause syntactic parsing errors and, consequently, errors in extracted information. A request, particularly rich in such errors, is given below. Unable to identify sentence-ending periods, Stanford syntactic parser has difficulty parsing this request.

(2) chromosome 3 found in the bloods between the father and son,, would this mean that my son,s blood is not the same as mine,, i was told it was all about learning problems,, but i am worried that there is more involved,, can you send me a chart or something describing ch 3 is all about and too what area of the body its being tested on espically a 9 year old child , thanks

In this paper, we present a method for detecting and correcting orthographical errors in consumer health questions. At this time, we do not attempt to correct grammatical errors. The method follows a pipeline architecture and consists of several modules: a) a pre-processing module that specifically focuses on errors involving punctuation and numbers, b) a misspelling detection module that relies on an expanded English dictionary, c) a spelling suggestion module that uses phonetic and orthographic distance, and d) a re-ranking module that uses a linear-weighted ensemble of several algorithms that score these suggestions. The scoring algorithms rely on orthographic, phonetic and contextual similarity as well as corpus frequency. We calculate contextual similarity using word embeddings, a recent natural language processing technique based on the distributional hypothesis. To evaluate our method, we developed a dataset that consists of 472 consumer health questions received by NLM, which we manually corrected for spelling errors. We compared the performance of our method against a strong baseline that relies on the ESpell algorithm, developed for PubMed query correction. Our results demonstrate the varying degrees of difficulty in correcting specific subtypes of spelling errors: while spelling errors in question elements most salient for question understanding (e.g., disorders) can be corrected to a large extent, real word spelling errors remain challenging.

2 Related Work

The approaches to correcting non-word spelling errors are reviewed by Kukich and Mitton. These approaches often rely on the availability of a comprehensive spelling dictionary and use edit distance and phonetic similarity between the misspelled word and the candidate suggestion. More recent approaches incorporate word frequency data collected from large corpora as well as contextual information. Flor and Futagi use word frequency data as well as orthographic and phonetic similarity, and re-rank the candidate suggestions using context, achieving very good correction results in correcting misspellings in non-native student essays. They report local error density (misspelling of adjacent words) and competition among inflectional variants as the main sources of their errors.

Real word spelling errors (also called malapropisms) are more challenging, because it is impossible to detect such errors in isolation. Thus, they often go undetected by spell checkers. Methods for real word error correction have used semantic information from lexical resources or relied on machine learning techniques and language models, all essentially taking some form of contextual information into account. The first approach is based on the hypothesis that the more distant a word is semantically from the other words in a text, the more likely it is a real word error. WordNet has often been used for calculating word distance. Machine learning techniques often use pre-defined confusion sets (e.g., \{their,there\}, \{principal,principal\}) and attempt to learn the typical context for each member using features from adjacent words. Language model-based approaches rely on n-gram probabilities generally drawn from web-scale data, such as
Google Web 1T dataset\(^1\). Syntactic and distributional information has also been used for this task\(^{[18]}\). While syntactic information based on parse features proved useful, the contribution of distributional information based on word cooccurrence was found to be limited.

In the biomedical domain, Crowell \textit{et al.}\(^{[19]}\) use a frequency-based technique to improve on existing spelling correction tools, ASpell and GSpell, for non-word spelling correction of consumer health information queries. They use MedlinePlus queries to generate word-frequency statistics. Their results show the significant contribution of frequency-based re-ranking for health information retrieval. Wilbur \textit{et al.}\(^{[7]}\) focus on spelling correction in the PubMed search engine. Their approach is based on the noisy channel model and makes use of statistics harvested from the user logs to estimate the probabilities of different types of edits that lead to misspellings. Word frequency counts in the PubMed database are used for computing prior probabilities. They apply different constraints based on the edit distance between the misspelling and the candidate suggestion. Ruch \textit{et al.}\(^{[20]}\) use the syntactic and semantic context to improve correction accuracy in clinical records. A simple, edit distance-based correction strategy is augmented with ranking via morpho-syntactic and word sense disambiguation. A named-entity extractor is used to avoid correcting physician and patient names. Patrick \textit{et al.}\(^{[21]}\) use a combination of a rule-based suggestion generation system and context-based and frequency-based ranking algorithm for spelling correction in clinical notes. Context-based ranking uses a trigram language model with the 3-word window around the misspelling. Each ranking method achieves the best result on one of the two test corpora. They also report inflectional variants as a challenge in spelling correction.

Word embeddings (also known as context-predicting models or neural language models\(^{[6,22]}\)) are a recent development in distributional semantics research, where the paradigm is to use vectors to represent the contexts that a word appears in and to apply vector algebra techniques to measure similarity between word vectors. In contrast to the more traditional distributional methods that rely on word counts, context-predicting models cast the problem as a supervised learning task and try to maximize the probability of the contexts that the word is observed in the corpus. The technique has been applied successfully to various word similarity tasks (e.g., semantic relatedness, analogy).

\section{3 Methods}

We begin this section by discussing our spelling correction dataset. We then explain the modules that comprise our spelling error detection and correction pipeline.

\subsection{3.1 Dataset}

We manually annotated and corrected 472 health-related questions posed to NLM by consumers for orthographic and punctuation errors. We considered the following types of errors in annotation:

- **NON-WORD**: the misspelled word does not appear in the dictionary (e.g., \textit{physians} for \textit{physicians})
- **REAL-WORD**: the misspelled word appears in the dictionary (e.g., \textit{leave} for \textit{live})
- **PUNCTUATION**: a spelling error caused by absence of punctuation or a spurious punctuation (e.g., \textit{I've} for \textit{I've})

\footnote{http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13}
• **TO-MERGE:** a word break error, where a spurious space is introduced to a word (e.g., *during* for *dur ing*)

• **TO-SPLIT:** a word break error, where two adjacent words run together (e.g., *knowabout* for *know about*)

In addition, we annotated whether the error occurred in a *focus* element of the question (**IMPORTANT-FOCUS**) or whether it was important for extracting the semantic frame (**IMPORTANT-FRAME**). All errors annotated as **IMPORTANT-FOCUS** can also be considered **IMPORTANT-FRAME**, since one major element of a frame is the disease in focus. A misspelled word may be annotated with multiple types of errors. For example, *onset* misspelled as *on set* is a case of both **TO-MERGE** and **REAL-WORD** errors. In Example 1, *serotonin* misspelled as *seretona* is annotated as **IMPORTANT-FOCUS**, as well, since *serotonin syndrome* is the focal disease in the question.

The annotation was carried out by one of the authors of this paper (MF). 1008 spelling errors were annotated on a total of 1075 tokens. Of these 1008 errors, 39 were labeled as **IMPORTANT-FRAME** and 96 as **IMPORTANT-FOCUS**. The distribution of error types annotated in the dataset is given in Table 1.

<table>
<thead>
<tr>
<th>Spelling error type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>NON-WORD</td>
<td>436</td>
</tr>
<tr>
<td>REAL-WORD</td>
<td>154</td>
</tr>
<tr>
<td>PUNCTUATION</td>
<td>58</td>
</tr>
<tr>
<td>TO-MERGE</td>
<td>45</td>
</tr>
<tr>
<td>TO-SPLIT</td>
<td>315</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>1008</strong></td>
</tr>
</tbody>
</table>

### Table 1: Distribution of spelling error types

3.2 **Pipeline**

The spelling error detection and correction pipeline, illustrated in Figure 1, consists of four modules. The *preprocessor* uses simple heuristics to correct punctuation and splitting errors that occur frequently and are difficult to correct using the methods applied to non-word or real word errors. For example, consumer health questions contain many contractions (e.g., *i’m*, *there’s*) and they are often spelled without the apostrophe. These are usually short words and most systems do not attempt to correct them, since the number of candidates based on edit distance and phonetic similarity is generally high. Therefore, we simply substitute such errors in preprocessing. For this purpose, we used Wikipedia’s list of English contractions[^2]. We also correct some informal expressions, such as *plz* for *please* and *u* for *you*. Additionally, we attempt to correct punctuation errors that could cause downstream tokenization or parsing errors, such as a punctuation without a space following it. In the example below, *male.reading* is taken as a single token, which would lead the system to miss the important information about the patient (*male*).

(3) *hi im a support worker for a 46 yr old autistic male.reading your info provided about wolfram syndrome ,he matches alot of the symptoms.*

Finally, we split tokens with leading or trailing digits, when we recognize that such a split would eliminate the spelling error (e.g., 4.5. Doctors for 4.5. Doctors, but not 2nd for 2nd).

The second step is to detect misspellings. As most other spelling error detection systems, we use a simple dictionary lookup. Our dictionary is based on the comprehensive English dictionary that is distributed with the Jazzy spell checker\(^3\), an open-source Java implementation based on ASpell. We expanded this dictionary with tokens extracted from the UMLS\(^23\), resulting in a dictionary of 450K tokens, including inflectional variants. Needless to say, misspelling detection step skips real word errors, which are both detected and corrected in the subsequent steps.

In the third step, we generate phonetic and edit distance-based spelling suggestions. We ensure that each suggestion is a valid word with dictionary lookup. For non-word spelling errors, tokens of length two or less, and for real word errors, tokens of length three or less, are skipped, unless they can be merged with the next token to form a valid word. Phonetic suggestions are obtained using the Double Metaphone algorithm\(^24\). We also compute suggestions using Levenshtein distance. The maximum number of edits is taken as the minimum of the half length of the token and 5.

The next step is to rank all generated suggestions using a set of algorithms. Three of these address the orthographic similarity between the misspelling and the suggestion:

- **Token similarity**: This measure is based on the cost of converting the misspelling to the suggestion, in terms of the number of edits required. Deletion and insertion operations have a higher cost than transposition and a lower cost than substitution. For example, the token similarity scores between the misspelling *dianosed* and the suggestions *diagnosed* and *deionized* are 0.91 and 0.61, respectively. Splitting the misspelling incurs an additional penalty.

- **Phonetic similarity**: Similar to token similarity, it uses the phonetic representation of the misspelling and the suggestion. The phonetic similarity between *dianosed* and *diagnosed* is 1.0, while it is 0.91 between *diagnosed* and *diagnose*.

\(^3\)http://jazzy.sourceforge.net/
• **Leading/trailing character overlap**: This measure calculates the overlap between the misspelling and the suggestion in terms of the number of matching characters at the beginning and the end. For example, the overlap score between *dianosed* and the word *diagnosed* is \((3+5)/9=0.89\), where the denominator corresponds to the length of the longer token.

In addition to these similarity measures, we score the suggestions by their frequency in consumer health-related corpora. To create the word frequency list, we used health-related articles in MedlinePlus Medical Encyclopedia\(^4\), MedlinePlus Drugs\(^5\), Genetics Home Reference\(^6\), Genetic and Rare Diseases (GARD) frequently asked questions\(^7\), NHLBI Health Topics\(^8\), NINDS Disorders\(^9\), and NIH Senior Health\(^10\), resources targeted at the public that we use for answering health-related questions. The number of articles from each resource are given in Table 2. We ignore numbers, URLs, and email addresses. The resulting word count list consists of more than 50K words. The corpus frequency score of a suggestion is calculated as

\[
freq\_score(s) = \ln(C_s/N)/\ln(C_{\text{max}}(w)/N)
\]

where \(C_s\) is the frequency of the suggestion, \(C_{\text{max}}(w)\) the maximum frequency count and \(N\) the number of tokens in the corpus, the score essentially corresponding to normalized unigram probability. Among the suggestions for *dianosed*, the score for *diagnose* is highest with 0.58; *deionized*, on the other hand, has a score of 0.

Finally, we calculate a contextual similarity score using word embeddings, taking context around the token into account. For this purpose, we used the same corpora that we used to calculate word frequency counts. We used the *word2vec* toolkit\(^11\) with the CBOW (continuous bag-of-words) model and hierarchical softmax options with window size of 5 to generate word vectors of 200 dimensions. CBOW, unlike the traditional bag-of-words model, uses the continuous distributed representation of the context. Hierarchical softmax is a computationally efficient way to estimate the overall probability distribution. To calculate the contextual

---

\(^6\)http://ghr.nlm.nih.gov/
\(^7\)http://rarediseases.info.nih.gov/gard
\(^8\)http://www.nhlbi.nih.gov/health/health-topics
\(^9\)http://www.ninds.nih.gov/disorders/disorder_index.htm
\(^10\)http://nihseniorhealth.gov/
\(^11\)https://code.google.com/p/word2vec/
similarity score for the suggestion, we compute the context vector by averaging the vectors associated with the words in a predefined window around the token under consideration and then compute the cosine similarity between the context vector and the suggestion vector. If there is no word vector corresponding to the suggestion or if the resulting cosine similarity is less than 0, we take the contextual similarity as 0. As context, we use two tokens before and after the token in consideration. If a token in the context does not have a corresponding vector, we move to the next token in the same direction. The contextual similarity score computed in this way is 0.69 for diagnosed for the context given in Example 1 and 0.51 for diagnose.

Scores calculated by these methods are then combined using a linear-weighted ensemble, similar to Flor and Futagi\cite{12}, taking the suggestion with the highest score as the correct spelling. Using the training set, we empirically determined the best weights to be 0.6 for orthographic similarity, 0.15 for contextual similarity, and 0.25 for frequency score.

For real word spelling errors, the methodology is essentially the same. We mentioned above that we skip valid tokens with three or less characters, unless they can be merged with the next token to form another valid token. In addition, for a valid token to be considered a real word error, we stipulate that: a) the difference between the contextual similarity of the suggestion and that of the original token be greater than a threshold value (taken as 0.2 in our experiments), b) the Levenshtein distance between the suggestion and the original token be equal to or less than 2, and c) the suggestion not be an inflectional variant of the original token and vice versa.

3.3 Evaluation

We used 372 questions for training and the remaining 100 for testing. As an informed baseline, we used the ESpell algorithm\cite{7}, designed for biomedical queries and used in the PubMed search engine. For non-word error correction, we used the Jazzy spell checker to filter out correctly spelled words for both ESpell and our ensemble method. For real-word correction, no such filtering was performed. We assessed the effect of preprocessing and each of the similarity measures on the system performance. We used precision, recall, and F1 score as the evaluation metrics.

4 Results and Discussion

The evaluation results obtained with the system are given in Table 3. Considering the non-word spelling errors only, the baseline, ESpell with non-word tokens as input, yields a F1 score of 0.29, whereas the best weighted combination of ranking methods yields a score of 0.61. The preprocessing module is able to correct a significant number of spelling errors, providing a solid foundation which the ranking methods improve on. Orthographic similarity seems to contribute most to the prediction of the correct spellings, providing a 12% increase in F1 score alone. Ignoring orthographic similarity and using only contextual similarity or corpus frequency diminishes the results. On the other hand, each of these methods complement orthographic similarity and together yield better precision and recall, pushing the overall F1 score to 0.61. When the system attempts to correct real word errors as well, the baseline, in this case ESpell without any filtering step, yields a F1 score of 0.25. This score is improved by more than two-fold to 0.58, when we use the setup that yields best results for non-word spelling errors with additional constraints. When we take into account only the spelling errors that were deemed to be important for focus or frame extraction, we obtain a F1 score of 0.76, in comparison to the baseline score of 0.57.

The corrections made with preprocessing may seem unimportant; however, we believe these corrections...
Table 3: Evaluation results

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F_1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-word only</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESpell (with filtering)</td>
<td>0.53</td>
<td>0.20</td>
<td>0.29</td>
</tr>
<tr>
<td>Preprocessing only</td>
<td>0.94</td>
<td>0.33</td>
<td>0.49</td>
</tr>
<tr>
<td>W/ Orthographic similarity</td>
<td>0.57</td>
<td>0.52</td>
<td>0.55</td>
</tr>
<tr>
<td>W/ Corpus frequency</td>
<td>0.45</td>
<td>0.41</td>
<td>0.43</td>
</tr>
<tr>
<td>W/ Context similarity</td>
<td>0.42</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>ALL</td>
<td>0.64</td>
<td>0.58</td>
<td>0.61</td>
</tr>
<tr>
<td><strong>Real-word included</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESpell</td>
<td>0.23</td>
<td>0.26</td>
<td>0.25</td>
</tr>
<tr>
<td>ALL</td>
<td>0.57</td>
<td>0.59</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>Important for focus/frame only</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESpell</td>
<td>0.58</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>ALL</td>
<td>0.83</td>
<td>0.70</td>
<td>0.76</td>
</tr>
</tbody>
</table>

will contribute to better tokenization and parsing, ultimately leading to better performance in information extraction. The results confirm that orthographic similarity is the best predictor of correct spellings. In general, edit distance of 2 or 3 is considered sufficient for successful spelling correction, based on the observation of Damerau\cite{10} that 80% of spelling errors are caused by a single edit. Our dataset provides evidence that spelling correction in consumer health questions may require considering suggestions with a higher edit distance (for example, in Example 1, *seretona* has an edit distance of 3 from its correct spelling, *serotonin*). This may be partly due to the fact that NLM receives questions from non-native speakers and partly due to the complexity of spelling of long medical terms.

We find that corpus frequency and contextual similarity contribute to spelling correction, though not as significantly as orthographic similarity. This may be partly due to the coverage of the corpora we considered. Their total size (approximately 5.8M tokens) is small, relative to the much larger corpora often employed in spelling correction. We experimented with larger corpora, such as one based on Wikipedia articles; however, we found that they did not make a significant difference, possibly because they are better written than health questions and are not health information-specific. To calculate contextual similarity, we use word embeddings, which, to our knowledge, is the first application of this technique to spelling correction. While the performance improvement due to word embeddings was relatively small, we believe it is a promising avenue to explore further. In its application to spelling correction task, one open question is how to best calculate contextual similarity using word vectors. We experimented with several measures and found that cosine similarity between the spelling suggestion and the average context vector yielded the best result; however, other measures (e.g., the maximum cosine similarity between the suggestion and a vector for any of the words in context) or a combination thereof could prove more beneficial.

The system, not unexpectedly, has more difficulty with real word spelling errors than with non-word errors. Our pipeline for real word errors differs little from that for non-word errors; it only incorporates several additional constraints. The results indicate that a more nuanced approach, probably with a different weighting scheme, may be necessary. It is worth noting that most systems focusing solely on real word spelling errors build on specific confusion sets\cite{15,18} or datasets to which real word errors are artificially introduced\cite{13,16},
and assume that the context around the error contains no spelling errors. We did not make such assumptions for the real word errors in our somewhat noisy dataset, and thus a fair comparison may be difficult.

Using a comprehensive, well-curated dictionary is critical for identifying misspellings. To have a better coverage of the biomedical domain, we extended an existing general English dictionary automatically with term tokens from UMLS, resulting in a comprehensive, albeit somewhat noisy, dictionary. Some errors were due to the noise in the dictionary. For example, the token Peyronies was simply skipped, since the word existed in the dictionary, although its canonical form, and the correct spelling in our dataset, is Peyronie’s.

Misspelling of adjacent words is a well-recognized problem in spelling correction and we observed errors due to this problem, as well. In Example 1, we are able to correct on set to onset based on contextual similarity when the following token is corrected from deminita to dementia; otherwise, we get a recall error.

5 Conclusion

We presented a spelling correction system for consumer health questions that takes into account orthographic features as well as corpus frequency information and contextual similarity. We did not attempt to use any semantic information (e.g., UMLS) beyond that obtained in an unsupervised manner from corpora. Our results show that all the components contribute to non-word spelling error correction, while real word errors remain challenging. Since the system is designed as part of a larger information extraction system, it is encouraging that it is able to recognize and correct the majority of spelling errors most relevant to information extraction.

The future work involves using more suitable corpora as the basis of frequency counts and word embeddings. We believe that a corpus that specifically focuses on text from online health forums would be more appropriate, since the questions we receive are probably closest to such text. We are currently incorporating such a corpus into our system. We are also planning to explore ways to better rank the similarity scores, probably in a machine learning framework. Another possible direction is to explore joint spelling correction, since misspelling of adjacent words is common in consumer health questions.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References


A Study of Concept Extraction Across Different Types of Clinical Notes

Youngjun Kim MS¹, Ellen Riloff, PhD¹, John F. Hurdle, MD, PhD²
¹School of Computing; ²Department of Biomedical Informatics,
University of Utah, Salt Lake City, UT

Abstract
Our research investigates methods for creating effective concept extractors for specialty clinical notes. First, we present three new “specialty area” datasets consisting of Cardiology, Neurology, and Orthopedics clinical notes manually annotated with medical concepts. We analyze the medical concepts in each dataset and compare with the widely used i2b2 2010 corpus. Second, we create several types of concept extraction models and examine the effects of training supervised learners with specialty area data versus i2b2 data. We find substantial differences in performance across the datasets, and obtain the best results for all three specialty areas by training with both i2b2 and specialty data. Third, we explore strategies to improve concept extraction on specialty notes with ensemble methods. We compare two types of ensemble methods (Voting/Stacking) and a domain adaptation model, and show that a Stacked ensemble of classifiers trained with i2b2 and specialty data yields the best performance.

Introduction
Medical notes provide detail on patient encounters. Written by clinicians primarily for clinicians, they document (e.g., progress notes) or summarize (e.g., discharge summaries) patient care. They come in a variety of note types and are entered by health care professionals from varying backgrounds.

Information extraction from medical texts is a challenging problem of growing interest to both the natural language processing and medical informatics communities. Medical concept extraction (MCE) is one such task, which seeks to identify specific types of information such as medical problems, treatments, and tests. Previous research on this task has primarily focused on discharge summaries and progress notes [1-4], which we will refer to as broad medical texts because they describe a patient’s overall care and their content can cover a diverse set of topics cutting across many areas of medicine. Most publicly available corpora of clinical medical notes consist of broad medical texts (e.g., i2b2 Challenge Shared Tasks [5-10] and ShaRe/CLEF eHealth Shared Tasks [11-12]). There has been relatively little research on medical concept extraction for more specialized clinical texts. Studies focused on Radiology and Pathology reports are an important exception, but we would argue that they also cover a broad set of clinical conditions. Broad medical texts have the advantage of being relatively well formatted, and they typically follow general documentation standards. In contrast, specialty notes conform to varying documentation standards, with little overlap between specialties. Patterson and Hurdle [13] and Friedman et al. [14] demonstrated that clinicians in different clinical domains use specific sublanguages. Still, given the general nature of broad medical notes, we speculate that their content could enrich MCE systems targeted at specialized note types and our work offers a practical way forward for clinical information extraction despite the common use of sublanguages.

Our research investigates methods for creating medical concept extraction systems that will perform well on specialty area notes. For this research, we created three new text corpora consisting of medical notes from three specialty areas: Cardiology, Neurology, and Orthopedics. We present an analysis of how they differ in content (semantic concepts and formatting) from each other and from i2b2 medical notes. We then examine a variety of information extraction (IE) models, and evaluate their performance on all of these data sets. The contributions of our work are twofold. First, we investigate how well MCE models perform on specialty notes when trained on a broad medical corpus and then when trained on the same type of specialty data. When training with a comparable amount of annotated data, we find that training with specialty texts outperforms training with broad medical texts. However, we achieve better performance for all three specialty areas by using a combination of both broad medical i2b2 data and specialty area data for training.

Second, we explore Voting and Stacked Learning ensembles to combine multiple MCE models. The ensemble architecture can be beneficial in two ways: (1) it can exploit multiple models that use different extraction techniques, and (2) it can exploit multiple models trained with different types of data (in our case, some trained on broad medical notes and some trained on specialty notes). To our knowledge, this is the first work that combines broad medical components and specialty area components in a single ensemble for MCE. Our results show that a stacked ensemble consisting of both types of components achieves the best balance of precision and recall.
Background

Medical concept extraction has been the focus of several shared tasks, such as the i2b2 Challenge Shared Tasks and the ShARe/CLEF eHealth Shared Tasks [8, 11-12]. Our work uses the annotated data set provided for the 2010 i2b2 Challenge [8]. In this challenge, machine learning approaches [15-16] showed superior results over hand-crafted rule-based systems. de Bruijn et al. [15] incorporated syntactic, orthographic, lexical, and semantic information (from various medical knowledge databases) and their system performed best in the i2b2 concept extraction challenge task with 83.64% recall, 86.88% precision, and 85.23% F1 score. Jiang et al. [16] implemented an ensemble method to combine concept extraction models trained with local features and outputs from different knowledge databases. In 2013, Tang et al. [17] extended their work using clustering and distributional word representation features, achieving 84.31% recall, 87.38% precision, and an F1 score of 85.82% on the i2b2 test set.

Our work is closely related to the classic task of Named Entity Recognition. In both newswire and biomedical texts, many types of supervised learning and sequential tagging methods have been used to extract specific types of entities [18-22]. In Clinical NLP (Natural Language Processing), several systems have been developed to process medical notes or biomedical texts. MedLEE [23] has been applied to chest radiology reports, discharge summaries, and operative reports to extract and encode medical information. MPlus [24] was used to extract medical findings, diseases, and appliances from chest radiograph reports. LifeCode [25] was developed to extract demographic and clinical information on emergency medicine clinical specialty and radiology reports.

Ensemble methods that combine multiple classifiers have been widely used for many NLP tasks. Voting strategies [26-28] and statistical approaches including stacked generalization [29-30] have generally shown better performance than individual classifiers. Our work is also related to supervised domain adaptation, which can be applied when some labeled data for the target domain is available. Many algorithms for efficient domain adaptation have been proposed, and domain adaptation-based models have been shown to improve performance for some tasks when limited annotated data is available for the target domain [31-35].

Methods

Data Sets and Annotated Concepts

Our research starts with the medical concept extraction (MCE) task defined for the 2010 i2b2 Challenge [8]. This task involves extracting three types of medical concepts: Problems (e.g., diseases and symptoms), Treatments (e.g., medications and procedures), and Tests. The 2010 i2b2 corpus consists of 349 training documents and 477 test documents manually annotated by medical professionals. This test set contains 45,009 annotated medical concepts.

For our work, we created new text collections representing three specialized areas of medicine: Cardiology, Neurology, and Orthopedics. We annotated 200 clinical notes from the BLULab corpus\(^1\) for each specialty area. Each specialty data set consists of different subtypes of notes. Table 1 shows the five most prevalent subtypes in each specialty data set.

Table 1. Five most prevalent note subtypes in each specialty area data set

<table>
<thead>
<tr>
<th>Data</th>
<th>Note subtypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiology</td>
<td>Cardiology (surgery) discharge summary, Cardiology (surgery) consultation report, Cardiology operative report, Cardiology history and physical examination, Angio report</td>
</tr>
<tr>
<td>Neurology</td>
<td>Neurosurgery discharge summary, Neurosurgery transfer summary, Neurology consultation report, Neurology history and physical examination, Neurosurgery death summary</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>Orthopedic (surgery) operative report, Trauma discharge summary, Orthopedic (surgery) discharge summary, Orthopedic surgery transfer summary, Orthopedics consultation report</td>
</tr>
</tbody>
</table>

\(^1\) The BluLab corpus is a collection of de-identified clinical notes drawn from multiple clinical settings at the University of Pittsburgh. The dataset was available for research to investigators with local Institutional Review Board approval, but unfortunately the University of Pittsburgh has withdrawn the corpus for new studies. However interested researchers can collaborate with previously approved sites.
Two people with medical expertise manually annotated the specialty notes using the 2010 i2b2 Challenge guidelines. One annotator had previously annotated data for the official 2010 i2b2 Challenge data and the other annotator had equivalent medical knowledge. We measured their inter-annotator agreement on 50 documents annotated by both annotators during the pilot phase using Cohen’s kappa [36] and their IAA was \( \kappa = .67 \). Each of the annotators then labeled 100 new documents for each specialty area, producing a total of 600 annotated specialty area texts. These texts contain 17,783 annotated concepts for Cardiology, 11,019 concepts for Neurology, and 12,769 concepts for Orthopedics.

Table 2 shows the number of annotated concepts of each type in the i2b2 test data and our three specialty data sets, as well as the average number of concepts per document. For example, the Cardiology data contains 7,474 Problem concepts and the average number of Problem concepts per text is 37, which is similar to the i2b2 data (39). However, the Neurology and Orthopedics data sets contain only 25 Problem concepts per document, on average. For Treatment concepts, the Neurology notes contain fewer than the i2b2 data but the Orthopedics notes contain more. The prevalence of Test concepts varies greatly: the i2b2 and Cardiology texts have many Test concepts per document, but they are much less common in the Neurology notes (11 per text) and Orthopedics notes (6 per text).

Table 2. The numbers of concepts in each data set

<table>
<thead>
<tr>
<th>Categories</th>
<th>i2b2 Test</th>
<th>Cardiology</th>
<th>Neurology</th>
<th>Orthopedics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Average</td>
<td>Total</td>
<td>Average</td>
</tr>
<tr>
<td>Problem</td>
<td>18,550</td>
<td>39</td>
<td>7,474</td>
<td>37</td>
</tr>
<tr>
<td>Treatment</td>
<td>13,560</td>
<td>28</td>
<td>5,706</td>
<td>29</td>
</tr>
<tr>
<td>Test</td>
<td>12,899</td>
<td>27</td>
<td>4,603</td>
<td>23</td>
</tr>
<tr>
<td>All Concepts</td>
<td>45,009</td>
<td>94</td>
<td>17,783</td>
<td>89</td>
</tr>
<tr>
<td># Sentences</td>
<td>45,052</td>
<td>94</td>
<td>21,255</td>
<td>106</td>
</tr>
</tbody>
</table>

The last row of Table 2 compares the number of sentences in the data sets. The i2b2 test data contains 45,052 sentences (94 per file, on average). The Cardiology notes were generally longer with 106 sentences per text, while the Neurology and Orthopedics notes were generally shorter.

We also examined, qualitatively, the types of sections in each data set to gain more insight about content differences between specialist notes and the more general i2b2 notes. Table 3 shows the five most frequent section titles in each data set. Many section titles, such as ‘Hospital course’, are common across all of the data sets. However, we found section titles that are much more frequent in some types of specialty area notes. For example, sections related to ‘Procedures’ and ‘Operations’ occurred most frequently in Orthopedics notes. ‘Consultation’ sections were common in the Cardiology notes, but rare in the i2b2 notes.

Table 3. Five most frequent section titles in each data set

<table>
<thead>
<tr>
<th>Data</th>
<th>Section Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>i2b2 Test</td>
<td>Hospital course, History of present illness, Physical Examination, Past medical history, Allergies</td>
</tr>
<tr>
<td>Cardiology</td>
<td>Physical examination, Allergies, Past medical history, Social history, History of present illness</td>
</tr>
<tr>
<td>Neurology</td>
<td>Hospital course, Reason for admission, History of present illness, Discharge medications, Discharge instructions</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>Hospital course, Procedures, Discharge instructions, Description of Operation, Complications</td>
</tr>
</tbody>
</table>

Although some of the same section titles occur in both broad medical notes and specialty notes, their contents can differ. For example, in the sections titled ‘Procedures,’ Orthopedics notes typically contain more detailed information than discharge summaries. Figure 1 illustrates an Orthopedics note that is similar to the ones in our collection.
Figure 1. A sample Orthopedics note

<table>
<thead>
<tr>
<th>PREOPERATIVE DIAGNOSIS:</th>
<th>Achilles tendon rupture, left lower extremity.</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSTOPERATIVE DIAGNOSIS:</td>
<td>Achilles tendon rupture, left lower extremity.</td>
</tr>
<tr>
<td>PROCEDURE PERFORMED:</td>
<td>Primary repair left Achilles tendon.</td>
</tr>
<tr>
<td>ANESTHESIA:</td>
<td>General.</td>
</tr>
<tr>
<td>COMPLICATIONS:</td>
<td>None.</td>
</tr>
<tr>
<td>ESTIMATED BLOOD LOSS:</td>
<td>Minimal.</td>
</tr>
<tr>
<td>TOTAL Tourniquet Time:</td>
<td>40 minutes at 325 mmHg.</td>
</tr>
<tr>
<td>POSITION:</td>
<td>Prone.</td>
</tr>
<tr>
<td>HISTORY OF PRESENT ILLNESS:</td>
<td>The patient is a 26-year-old African-American male who states that he was stepping off a hill at work when he felt a sudden pop in the posterior aspect of his left leg. The patient was placed in posterior splint and followed up at ABC orthopedics for further care.</td>
</tr>
<tr>
<td>PROCEDURE:</td>
<td>After all potential complications, risks, as well as anticipated benefits of the above-named procedure were discussed at length with the patient, informed consent was obtained. The operative extremity was then confirmed with the patient, the operative surgeon, Department Of Anesthesia, and nursing staff. While in this hospital, the Department Of Anesthesia administered general anesthetic to the patient. The patient was then transferred to the operative table and placed in the prone position. All bony prominences were well padded at this time. A non-sterile tourniquet was placed on the left upper thigh of the patient, but not inflated at this time. Left lower extremity was sterilely prepped and draped in the usual sterile fashion. Once this was done, the left lower extremity was elevated and exsanguinated using an Esmarch and the tourniquet was inflated…</td>
</tr>
</tbody>
</table>

Information Extraction Models

We developed four types of information extraction models that use a diverse set of extraction techniques.

**Rules:** We created a simple set of rules by harvesting information from the annotated training data. First, for each word in the training data we computed Prob(concept | word) and Prob(category | word). Next, we selected words that had frequency $\geq 3$ and Prob(concept | word) $\geq .80$. For each selected word, we chose the category with the highest probability and created a rule (e.g., `diabetes` $\rightarrow$ `Problem`). Given a new text, we then found all words that matched a rule and labeled them as concepts using the category assigned by the rule. When two or more labeled words were contiguous, we treated them as a single concept. For multi-word concepts, we calculated the average Prob(category | word) across the words in the concept. The category with the highest average probability was assigned to the concept.

**MetaMap:** We used a well-known knowledge-based system, MetaMap [37], that assigns UMLS Metathesaurus semantic concepts to phrases. We identified UMLS semantic type identifiers, using the UMLS Semantic Lexicon, that covered the types of medical concepts required for our task. We only used the final mappings of MetaMap to avoid generating nested terms because the i2b2 guidelines do not permit nested concepts. Table 4 shows the semantic types that we used for concept extraction. We used MetaMap 2013v2 with the 2013AB NLM relaxed database.

<table>
<thead>
<tr>
<th>Category</th>
<th>MetaMap semantic types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem</td>
<td>acab, anab, bact, celf, cgab, chvf, dsyn, inpo, mobd, neop, nnon, orgm, patf, sosy</td>
</tr>
<tr>
<td>Treatment</td>
<td>antb, carb, horm, medd, nsba, opec, orch, phsu, sbst, strd, topp, vita</td>
</tr>
<tr>
<td>Test</td>
<td>biof, bird, cell, chvs, diap, enzy, euka, lbpr, lbtr, mbtr, moft, phsf, tisu</td>
</tr>
</tbody>
</table>

2 Excerpted from http://www.mtsamples.com/
3 Refer to http://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt for the mapping between abbreviations and the full semantic type names.
SVM: We trained a multi-class Support Vector Machine (SVM) classifier with a linear kernel using the LIBLINEAR software package [38]. We applied the Stanford CoreNLP tool [39] to our data sets for tokenization and part-of-speech (POS) tagging. We defined features for each medical concept’s lexical string, POS tag, affix(es), orthographic features, and pairwise combinations of these features. We also extracted these features for the three words before and after the concept. To identify multi-word concepts, we reformatted the training examples with IOB tags (B: at the beginning, I: inside, or O: outside of a concept) and then trained the model to produce IOB labels as output. We trained a single SVM model to produce labels for all three concept types (Problem, Treatment, and Test).

CRF: We trained two types of sequential taggers using linear Conditional Random Fields (CRF) models [19]: models with forward transitions (CRF-fwd) and models with backward transitions by reversing the word sequence (CRF-rev) [40-41]. The CRF models used the same feature set as the SVM models. A single CRF model produces labels for Problem, Treatment, and Test.

We performed 10-fold cross validation on the i2b2 training set to optimize the parameters of the SVM and CRF classifiers for F1 score maximization. For the SVMs, we tuned the cost parameter (c = 0.1) of LIBLINEAR. For the CRFs, we used Wapiti [42], a simple and fast discriminative sequence labeling toolkit. We set the size of the interval for the stopping criterion to be $e = .001$. For regularization, $L_1$ and $L_2$ penalties were set to 0.005 and 0.4 respectively. These parameter settings were kept the same throughout all of our experiments.

Ensemble Methods

We explored two types of ensemble architectures that have performed well for other NLP tasks: Voting ensembles and Stacked Learning ensembles [29]. Each ensemble consists of a set of MCE components. The general architectures of the Voting and Stacked ensembles are described below. In the Results section, we present experimental results for ensembles consisting of different mixtures of component systems.

Voting Ensemble: This ensemble collects the phrases labeled by a set of MCE components and outputs all phrases that received at least three votes (i.e., were labeled by at least three components). In the case of overlapping phrases, we choose the one with the highest confidence, based on the normalized confidence scores of the MCE models. For each MCE model, each confidence score was divided by the highest score produced by that model for normalization.

Stacked Learning Ensemble: This ensemble consists of a set of MCE components as well as a meta-classifier, which is an SVM classifier trained on the predictions of the individual MCE models. To create training instances for a document, we first aggregated all of the concept predictions into sets of unique predictions. For example, one aggregated prediction set might indicate that an instance of “acute renal failure” was labeled as a Problem by models $M_1$, $M_3$, and $M_4$. Each concept predicted by an MCE model was then compared with all concepts predicted by the other MCE models. For each pair of concepts, the following eight matching criteria are applied to create binary features:

- If the text spans match
- If the text spans partially match (any word overlap)
- If the text spans match and concept types match
- If the text spans partially match and the concept types match
- If the text spans have the same start position
- If the text spans have the same end position
- If one text span subsumes the other
- If one text spans is subsumed by the other

For an ensemble with $k$ MCE models, $k \times 8$ features are defined so that each matching function is replicated for each MCE model. Each feature indicates whether Concept$_1$ and Concept$_2$ satisfy a specific matching function, given that Concept$_1$ was produced by a specific model. In addition, features are defined that count how many different models produced a predicted concept, and features are defined for predictions produced by just a single model (indicating which model produced the predicted concept). In a previous study [43], this type of Stacked ensemble architecture achieved performance comparable to the state-of-the-art on the i2b2 test data with 83.4% recall, 87.9% precision, and 85.6% F1 score.

Results

We conducted an extensive set of experiments to evaluate the performance of each individual MCE model and Voting and Stacked Learning ensembles. We also experimented with models trained using the broad medical (i2b2) texts, using our specialty area texts, and using a mixture of both. We evaluated performance using the i2b2 test set as
well as our three sets of specialty area notes: Cardiology, Neurology, and Orthopedics. The specialty area models (Sp) were trained and evaluated using 10-fold cross validation on our specialty notes data. A labeled phrase was scored as correct if it was assigned the correct concept type and its text span exactly matched the gold standard text span, disregarding articles and possessive pronouns (e.g., “his”).

Performance of Individual MCE Models

Table 5 shows the performance of each MCE model based on Recall (Rec), Precision (Pr), and F1 score (F). The Rules (i2b2) row shows results for the simple rules harvested from the i2b2 training data. Not surprisingly, these rules performed better on the i2b2 test set than on the specialty notes, but the scores were low across the board. The Rules (Sp) row shows results (averaged during cross-validation) for the rules harvested from the training folds and evaluated on the test folds for the specialty area data. These rules also performed poorly. The MetaMap row shows similarly low scores for MetaMap on all data sets. One reason for its low performance is that the concept and phrase boundary definitions of MetaMap’s semantic categories are not perfectly aligned with i2b2’s concept definitions.

The machine learning classifiers performed substantially better. The SVM (i2b2) row shows results for the SVM model trained on i2b2 data, which produced an F1 score of 78.7% on the i2b2 test set but substantially lower F1 scores on the specialty datasets. The SVM (Sp) row shows results for the SVMs trained on specialty area data. Performance substantially improved on the Orthopedics notes (from 43.5% to 52.6% F1 score), but did not change much for the other specialty areas.

<table>
<thead>
<tr>
<th>Model</th>
<th>i2b2</th>
<th>Cardiology</th>
<th>Neurology</th>
<th>Orthopedics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rec</td>
<td>Pr</td>
<td>F</td>
<td>Rec</td>
</tr>
<tr>
<td>Rules (i2b2)</td>
<td>38.5</td>
<td>48.4</td>
<td>42.9</td>
<td>33.1</td>
</tr>
<tr>
<td>Rules (Sp)</td>
<td></td>
<td></td>
<td></td>
<td>32.6</td>
</tr>
<tr>
<td>MetaMap</td>
<td>36.0</td>
<td>47.3</td>
<td>40.9</td>
<td>31.1</td>
</tr>
<tr>
<td>SVM (i2b2)</td>
<td>80.6</td>
<td>76.9</td>
<td>78.7</td>
<td>64.5</td>
</tr>
<tr>
<td>SVM (Sp)</td>
<td>65.5</td>
<td>59.4</td>
<td>62.3</td>
<td>60.2</td>
</tr>
<tr>
<td>CRF-fwd (i2b2)</td>
<td>81.4</td>
<td>86.1</td>
<td>83.7</td>
<td>65.2</td>
</tr>
<tr>
<td>CRF-rev (i2b2)</td>
<td>82.3</td>
<td>86.4</td>
<td>84.3</td>
<td>65.8</td>
</tr>
<tr>
<td>CRF-rev (i2b2,180)</td>
<td>78.7</td>
<td>84.2</td>
<td>81.4</td>
<td>63.3</td>
</tr>
<tr>
<td>CRF-fwd (Sp)</td>
<td>63.8</td>
<td>69.3</td>
<td>66.4</td>
<td>59.2</td>
</tr>
<tr>
<td>CRF-rev (Sp)</td>
<td>65.2</td>
<td>69.1</td>
<td>67.1</td>
<td>60.5</td>
</tr>
<tr>
<td>CRF-rev (i2b2+Sp)</td>
<td>68.7</td>
<td>70.3</td>
<td>69.5</td>
<td>64.6</td>
</tr>
</tbody>
</table>

Both the CRF-fwd and CRF-rev models trained on i2b2 data performed better than the SVM models. Performance on the Cardiology and Neurology notes was similar when trained on specialty (Sp) data, but performance on the Orthopedics notes substantially improved. Since the i2b2 training data is much larger than our specialty area training data, we performed another experiment using only 180 randomly selected i2b2 training texts, to match the amount of specialty area training data (under 10-fold cross-validation, each fold trains with 180 documents). The performance of this model, shown in the CRF-rev (i2b2,180) row, is lower than when using all of the i2b2 training data. We can now see that training on specialty area data consistently performs better than training on i2b2 data when using comparable amounts of training data. The last row of Table 5 shows the results for training the CRF-rev model using all of the i2b2 training data as well as the specialty area training data. Performance improved for all three specialty areas by training with the combined data sets. The broad i2b2 data clearly provides added value. However, the F1 scores for the three specialty areas ranges from 60.9% to 69.5%, which is substantially lower than the 84.3% F1 score achieved for the i2b2 test set.
Performance of Voting and Stacked Ensembles

We also evaluated the performance of the Voting and Stacked ensemble architectures, which were populated with all five types of MCE components: Rules, MetaMap, SVM, CRF-fwd, and CRF-rev models. For both the Voting and Stacked architectures, we created three different types of ensembles: i2b2 ensembles consisting of MCE models trained on the i2b2 data, Sp ensembles consisting of MCE models trained on specialty data, and i2b2+Sp ensembles consisting of MCE models trained on i2b2 data and MCE models trained with specialty data. Consequently, the i2b2+Sp ensembles include nine different classifiers (two models each of Rules, SVM, CRF-fwd, CRF-rev, and one MetaMap model, because it does not use training data).

Table 6 shows the performance of these ensembles, as well as the EasyAdapt domain adaptation method [34], which we implemented as another point of comparison. For EasyAdapt, we used a CRF-rev classifier with the feature set augmented for broad medical (i2b2) notes as the source domain and specialty area notes as the target domain. For the sake of comparison, the first row of Table 6 displays again the results obtained for the best individual MCE model from Table 5, which was the CRF-rev classifier trained with both i2b2 and specialty data. Comparing the first two rows, we see that training a CRF-rev model with combined i2b2 and specialty area data outperforms the domain adaptation model on all three data sets.

For the Voting ensembles, the i2b2+Sp ensemble produced the best \( F_1 \) scores, but did not outperform the CRF-rev (i2b2+Sp) model. However, the Voting ensemble trained only on specialty notes (Sp) produced much higher precision than the CRF-rev model. A Voting ensemble appears to be an effective way to improve precision on specialty notes when a limited amount of annotated specialty data is available, although with some cost to recall.

For Stacked Learning, every Stacked ensemble outperformed its corresponding Voting ensemble. The best Stacked ensemble (i2b2+Sp) included MCE models trained on i2b2 data as well as MCE models trained on specialty data, producing slightly higher \( F_1 \) scores than the CRF-rev models for all three specialty areas. Using a paired t-test to measure statistical significance, the \( F_1 \) score performance of the i2b2+Sp Stacked ensemble is significantly better than EasyAdapt and all of the Voting ensembles at the \( p < .05 \) significance level, but not significantly better than the CRF-rev (i2b2+Sp) model. However, the results show that the Stacked ensemble produces higher precision than the CRF-rev model (70% → 75% for Cardiology; 67% → 72% for Neurology; 63% → 71% for Orthopedics), with correspondingly smaller decreases in recall (69% → 66% for Cardiology; 65% → 62% for Neurology; 59% → 55% for Orthopedics).

**Discussion and Analysis**

The main conclusion of our research is that models trained with a combination of broad medical data and specialty data consistently perform better than models trained on either type of data alone when the amount of specialty data is limited. In addition, we find that a Stacked ensemble consisting of a diverse set of MCE models using different types of extractors achieves overall performance comparable to the best individual classifier in our experiments, but
offers two advantages. First, the Stacked ensemble yields a recall/precision balance that favors precision, which may benefit applications that place a premium on high precision. Second, the Stacked ensemble can be easily augmented with additional components as new resources become available, because the meta-classifier automatically learns how to use them simply by re-training the meta-classifier component. In contrast, adding new components to Voting ensembles can require a change in voting strategies, and Voting ensembles do not provide a way to learn weights to optimally control the influence of different component models.

However, performance on all three types of specialty areas is much lower than performance on the broad medical (i2b2) texts. Clearly there is ample room for improvement for medical concept extraction from specialty area clinical notes and more work is needed on this topic. To better understand the strengths and weakness of our models, we manually inspected their output. We observed that our ensemble methods are particularly successful at identifying more accurate concept boundaries than the individual MCE models (e.g., identifying “severe chest pain” as a Problem concept instead of just “severe” or “chest pain”). We also analyzed the false negative errors by the CRF-rev models trained with i2b2 data and those trained with specialty data. Table 7 shows the results of this manual analysis, which were based on one test fold (20 notes) for each specialty area. The first row of Table 7 corresponds to errors due to unseen vocabulary. These concepts were misclassified when none of the words in a concept occurred in the training data. For example, the Cardiology concepts ‘thoracoscopy’ and ‘cardioplegia’ never appeared in the i2b2 training data. Unseen concepts accounted for roughly the same percentage of errors when training with i2b2 data or specialty data, but note that the i2b2 training set is roughly twice as large as each specialty area training set.

Table 7. False negatives errors by CRF-rev (i2b2) and CRF-rev (Sp) models

<table>
<thead>
<tr>
<th>Error types</th>
<th>Cardiology</th>
<th>Neurology</th>
<th>Orthopedics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i2b2</td>
<td>Sp</td>
<td>i2b2</td>
</tr>
<tr>
<td>All unseen</td>
<td>22 (5%)</td>
<td>27 (6%)</td>
<td>31 (6%)</td>
</tr>
<tr>
<td>At least one unseen word</td>
<td>138 (31%)</td>
<td>100 (21%)</td>
<td>186 (37%)</td>
</tr>
<tr>
<td>At least one word rarely seen</td>
<td>70 (16%)</td>
<td>81 (17%)</td>
<td>70 (14%)</td>
</tr>
<tr>
<td>All seen</td>
<td>213 (48%)</td>
<td>279 (56%)</td>
<td>211 (43%)</td>
</tr>
</tbody>
</table>

The second row of Table 7 corresponds to false negatives for concepts containing at least one seen word and one unseen word. We see more false negatives in this category for the models trained with i2b2 data than the models trained with specialty data. For example, for the Treatment concept ‘aortic crossclamping’, ‘crossclamping’ never appeared in the i2b2 training data but it did appear in the Cardiology training data. This type of error was most common in the Orthopedics data (51% of the errors), which suggests that the Orthopedics notes contain many vocabulary terms that are not present in the i2b2 data.

The third row of Table 7 corresponds to false negatives for concepts containing all seen words, but at least one rarely seen word (frequency <= 3). For example, in the Cardiology data, the concepts ‘psa data’ and ‘r-wave’ were not identified by the i2b2 trained model. The model trained with Cardiology data could not extract ‘nystatin’ and ‘oximeter’, even though they occurred (infrequently) in the Cardiology training data.

The last row of Table 7 corresponds to false negatives for concepts consisting entirely of words that occurred > 3 times in the training data. Many false negative errors fell into this category. Generally, there were more false negative errors of this type for the models trained with specialty data than those trained with i2b2 data, presumably because the vocabulary is more homogenous in the specialty areas so more words simply fall into the seen category.

Finally, we observed that many errors were due to incorrect phrase boundaries of medical concepts. For example, only the word “hepatitis” was labeled in the phrase “hepatitis c”. We also witnessed some tricky errors due to contextual differences in the words surrounding medical concepts. For example, a Treatment concept ‘lidocaine’ is often prescribed for usage on skin (“treated with lidocaine jelly for pain control”). However, in the Cardiology data, it is usually applied by infiltration (“Lidocaine 20 cc was infiltrated into the tissues”).

Conclusion

We analyzed the differences in content between broad medical and specialty area notes, confirming prior research showing that specialty notes exhibit sublanguage behavior that requires rethinking the use of NLP tools developed.
on broad medical notes. We found that even though the CRF-rev (i2b2+Sp) and Stacked ensemble produce similar F_1 scores, they exhibit different behaviors with respect to the underlying recall and precision of their output. Consequently, our results suggest that the CRF-rev (i2b2+Sp) model may be preferable for applications where recall is more important than precision, while the Stacked ensemble may be preferable for applications where precision is more important than recall. Interestingly, Orthopedics specialty notes exhibit the most unique language when compared to other specialty notes or to broad medical texts. When a limited amount of annotated specialty area data is available, our research shows that training concept extractors with both broad medical data and specialty area data produces MCE models that achieve better performance on specialty notes than training with either type of data alone. In addition, our research found that a Stacked ensemble with a mixture of MCE components, including different types of MCE models as well as models trained on different types of data, achieves good performance and offers some advantages over other approaches. However, we also observed that MCE performance on specialty texts is substantially lower than state-of-the-art performance on broad medical texts. A promising direction for future work is to explore semi-supervised methods to exploit larger collections of specialty area notes for training.

Acknowledgments

This research was supported in part by the National Science Foundation under grant IIS-1018314 and the National Library of Medicine under grant R01-LM010981. We thank Jennifer Thorne, RN and Jenifer Williams, RN for their annotation work, and Dr. Stéphane Meystre for his comments and feedback.

References

Supporting Multi-sourced Medication Information in i2b2

Jeffrey G. Klann, PhD1,2,3, Pascal B. Pfiffner, MD, PhD2,4, Marc D. Natter, MD, MS2,4, Emily Conner5, Paul Blazejewski4, Shawn N. Murphy, MD, PhD1,2,3, Kenneth D. Mandl, MD, MPH2,4

1Partners Healthcare, Boston, MA; 2Harvard Medical School, Boston, MA; 3Massachusetts General Hospital, Boston, MA; 4Boston Children's Hospital, Boston MA; 5Anotera, LLC

Abstract
Postmarketing drug surveillance is critical to assessing adverse events associated with medications, because prelaunch clinical trials frequently miss negative drug effects. The Informatics for Integrating Biology and the Bedside platform (i2b2) has been used effectively for this. However, previous work suffers from incomplete medical data present in electronic health record (EHR) systems. Here, we develop a system to integrate non-traditional data sources with EHR data: pharmacy dispensing information and patient-reported data. We implement and validate a toolset to gather medication data from a Pharmacy Benefit Manager network, import it into an i2b2 EHR repository using a standard data format, merge it with the EHR data, and present it to for annotation with results returned to i2b2. This toolkit is enabling studies on medication list data quality, adherence, and adverse event detection.

Introduction
Postmarketing drug surveillance is critical to assessing adverse events associated with medications, because prelaunch clinical trials frequently miss negative drug effects. Such clinical trials have limitations in population size, diversity, disease comorbidities, and concurrent medication use. [1] Incident Reporting Systems, such as FDA MedWatch, are used to manually report postmarketing adverse events. However, their utility is limited by underreporting and poor compliance by physicians. [2] The Informatics for Integrating Biology and the Bedside platform (i2b2) has been used effectively for postmarketing drug surveillance. [3,4] i2b2, the output of an NIH National Center for Biomedical Computing grant, is an open source medical data analytics platform which is in use at over 100 sites nationwide that have data on over 90 million patients.

However, incomplete medical data is rampant, and because i2b2 relies on the "ambient" data generated by the electronic health record systems, it suffers as well. Medication histories are frequently poorly maintained, which is not aided by the significant fragmentation of care across sites (increasing the likelihood of incomplete data at any given site). [5–7]

We hypothesize that it is possible to improve the medication list by enlisting nontraditional data sources. We have begun to explore two sources thus far.

First, medication dispensing information, recorded by pharmacies when patients pick up medications, is fueled by the pharmacy supply chain. The pharmacy supply chain already has rather advanced data networks, traditionally used for payments, claim adjudications, and inventory management. [8] These data can be used as a source for postmarket drug surveillance. [9]

Second, the importance of patient reported data is increasingly becoming recognized. [10,11] Such data have been found to be timely and reliable, and patients are frequently willing to share and participate in their healthcare. [12,13] Along these lines, the Patient Centered Outcomes Research Institute recognizes and emphasizes patient centered research. They are leading “a national movement that meaningfully involves patients and incorporates their voices in producing trusted, evidence-based information, promoting better decisions and ultimately better health for all.” [14]

We previously developed the i2b2 Standards-based Extract, Transform, and Load (SETL) tool, which integrates i2b2 data with data in a standardized format, the Consolidated Clinical Document Architecture (C-CDA). [15] This work was timely because an embodiment of C-CDA is required by Meaningful Use and is therefore becoming widely available. [16,17] Our previous work supports import of medication prescribing events in C-CDA, which originate from the EHR. However, C-CDA can also represent many types of medication events with high granularity, including prescribing, billing, and dispensing events.

In this manuscript, we leverage our previous work on the SETL tool to develop a patient-centered medication annotation platform for i2b2 that can import granular medication information from a C-CDA source, combine it with
existing EHR data stored in i2b2, and enable multi-source patient medication annotation via a mobile app. We validate this platform by capturing pharmacy-benefit-manager-derived (PBM) data from a commercial PBM that is a major source of pharmacy dispensing data into i2b2 on an on-demand basis. We transfer PBM data into the i2b2 clinical instance at Boston Children’s Hospital alongside EHR data, and validate the app’s ability to interpret these data for patient action. This platform will enable forthcoming research in medication data quality, adherence, and adverse event detection. Presently, researchers are preparing to use this platform to study medication adherence by assessing the discordance between the clinical medication list and a patient-generated list.

Methods

Design

C-CDA. The C-CDA is a rich but complex XML document language for healthcare derived from the Health Level Seven (HL7) Reference Information Model. [16] It is elegant but so expressive that differences of opinion arise on the best way to represent healthcare events. A great deal of work has gone into defining Meaningful Use templates for the Continuity of Care Documents required by that program. However, no such templates exist for the differentiation of medication prescribing, dispensing, billing, and patient-provided medication data. We augmented the Meaningful Use templates with existing C-CDA templates to support the complete scope of medical data required for this study.

HL7v2 and PBM. PBM data is conveyed in record-based HL7 version 2 format (HL7v2), which is a declining but still predominant format for the exchange of healthcare data. Although used as foundation for most older interoperable healthcare systems, its record-based format is recognized as limited and many data sources that use HL7v2 add custom extensions to the format in order to achieve expressiveness; this can be a hindrance to interoperability. We developed a mapping from the PBM elements of HL7v2 to C-CDA.

i2b2 and Patient Data Objects. The i2b2 software is compartmentalized into functional units known as cells. These cells, which collectively make up the i2b2 hive, interact through web service calls. Clinical data is exchanged with these web services in the i2b2 Patient Data Object format, which is a denormalized data structure that reflects the underlying star schema design of the i2b2 database. [18] The atom of a PDO is the observation. Fact observations record clinical information associated with the patient, such as a medication he/she is taking. Modifier observations are linked to fact observations and provide additional information, such as medication route, dose, and information source (e.g., medication list, dispensing event, or patient report). The set of possible observations are defined as trees of facts and modifiers through i2b2 ontology services. Although some standard ontology trees are emerging as i2b2 is increasingly used for interoperable networks (for which a standardized ontological information model must be defined) [19], i2b2’s general position has been to leave development of the ontology trees to individual use cases, thus maximizing flexibility. We expanded our previous C-CDA to PDO mapping to include the more complex C-CDA documents used here.

System description

The architecture of the platform is outlined in Figure 1. This necessitated development of three new software components:

- A new i2b2 cell, the i2me2 cell. It includes two main functions, which are similar conceptually to previous work, but with important differences:
  - The first function is similar to the SETL cell previously described: transform data in a standard format (C-CDA) into i2b2 PDO format. [15] Like the SETL cell, the i2me2 C-CDA transformation functionality relies on the Model Driven Message Interoperability (MDMI) toolkit, and it loads data into the i2b2 EHR repository using the i2b2 Clinical Research Chart (CRC) Loader web service. Unlike the SETL cell, the i2me2 cell supports the more complex medication C-CDAs developed here.
  - Second, the i2me2 cell manages authentication, authorization, and communication with the medication annotation app. Authentication and authorization are accomplished separately from i2b2 in order to accommodate the enhanced granularity of access control and auditing required.
This additional layer of authorization augments i2b2's access controls, which offer access levels (such as full access or aggregate counts only) only across the entire repository, due to its cohort study focus. The i2me2 authorization schema enables subjects full access to only their own data, augmented by personally identifying information not available to i2b2. This is accomplished through an Identity and Demographics Module (IDM) within i2me2. The cell transmits a subject-specific token to the annotation app that is first used to request limited demographic data (e.g. name) to confirm the identity of the subject. This token is then passed back to i2me2 with a request for the subject's medication data. Upon receipt of this token, the i2me2 cell obtains medication data from i2b2 and transmits it to the app, possibly requesting medication information from RxConnect if PBM data has not already been imported into the i2b2 EHR repository. Finally, following use of the app, the updated medication list is transmitted from the app to i2me2 in PDO format. This is added to the i2b2 repository by i2me2 using the CRC Loader service. This follows a similar approach to the previously described i2b2-SMART cell, which uses the SMART RDF format to provide medical apps with subject-specific i2b2 data. [20] Because the i2b2-SMART cell presently supports neither bi-directional communication nor differentiation of medication event type (e.g., prescribing or dispensing), we currently use this more direct PDO approach.

A software component, RxConnect, provides a synchronous interface to the PBM web service. RxConnect receives patient demographic information from the IDM module via i2me2, which is used to identify a subject in the PBM patient index. The PBM then asynchronously returns medication dispense and claims information for the respective subject. RxConnect receives the HL7v2 response message from the PBM, parsing this using the HAPI open-source toolkit (http://hl7api.sourceforge.net). The returned message is processed, filtered of personally identifying information, and then converted to C-CDA via MDMI.

- A patient-facing application enabling patient input into the i2b2 medication data model. The app consumes an i2b2 PDO for a single subject that can reflect medication data from multiple sources. A list of de-duplicated medications is generated for presentation, one medication at a time, to the subject performing medication self-validation. Using a two-step process, the subject indicates if and how often a medication is taken. Missing medications may also be added. The patient-annotated medication list is then added back to the i2b2 repository to as patient reported data, which exists alongside (and not in replacement of) other sources of medication data in the repository. This multi-sourced data is then available to researchers for analyses.

The data conversions from the record-based HL7v2 format to the document based C-CDA format were developed with the MDMI toolkit, which is an open source project in the Open Health Tools (OHT) consortium that seeks to make interoperability a commodity. [21] MDMI provides a graphical data-mapping tool that supports rapid design and iterative development and can move data in any arbitrary data format to and from the MDMI model, known as the Referent Index. This Referent Index is an evolving standard that attempts to represent all possible healthcare data using a common semantic structure based elements identified in the C-CDA specification. By divorcing semantics from data structure, MDMI provides a unique approach to move data between different types of formats (in this case, record based, document-based, and denormalized).

These software components enable nontraditional data sources (pharmacy dispensing data and patient annotation information) to be added to an i2b2 EHR repository of medication data in five steps (shown in Figure 1).

1. Subject information is requested from the PBM network, which asynchronously provides dispensing and insurance claims information for a particular subject in HL7v2 format. This may be scheduled or on-demand.
2. RxConnect processes the message, converts it to C-CDA, and sends it to the i2me2 cell.
3. The i2me2 cell converts the incoming message from C-CDA to i2b2 PDO format as needed, and loads this data into the i2b2 repository alongside the pre-existing EHR data.
4. When a subject is ready to annotate their medication information, the i2me2 cell requests a combined PDO containing both dispensing data (from the PBM) and prescribing data (from the i2b2 EHR repository). The PDO is then forwarded securely to the app.
5. Subjects use the app to construct a current medication list that is added back to the i2b2 EHR repository as patient reported data to exist alongside (and not in replacement of) other sources of medication data in the repository.

749
The output of the five steps is a rich database of medication information containing prescribing, dispensing, and patient-reported medication data, as well as other i2b2 data sources including EHR and disease registry medication data. This will enable more accurate adverse drug event research. All data in the i2b2 repository can be studied via the i2b2 web client query tool, which supports dimensional data queries using the ontology hierarchies.

This patient annotation platform is not intended for clinical use at this time – the required quality testing would be a costly and time-consuming endeavor that is unnecessary for research use. The platform is research-focused and its usage will vary depending on the study question. Patients must be recruited for a particular IRB-approved study question before the five steps above can be performed.

Results

![Diagram of the i2me2 architecture](image)

**Figure 1**: The “i2me2” architecture, outfitted to combine i2b2 EHR medication data with PBM dispensing and claims data, as well as patient-reported medication information. Data flows from the PBM into i2b2 (steps 1-3), then to the subject for annotating corrections, and finally back to the i2b2 data repository. (Steps 4-5) The data are transformed from HL7v2 to the standard C-CDA to i2b2 Patient Data Objects.

Implementation

In this proof-of-principle, we implemented the architecture in Figure 1 at Boston Children’s Hospital and obtained 11 messages from the PBM corresponding to 11 synthetic patients intended for testing. We used these cases to develop our HL7v2 to C-CDA to PDO mappings, using the graphical MDMI Map Editor.
Our approach to constructing valid C-CDA and PDO messages involved a number of considerations due to complexities in data transformation.

First, the meaning of the HL7v2 medication order segment (i.e. dispensing vs. payment event) is determined by a custom element defined by the PBM. Although the MDMI specification is designed to support such overloaded fields through “linked data” functionality, the current MDMI transformation engine did not provide this functionality, as MDMI’s focus is declarative transforms. Declarative transforms encourage map developers to adopt straightforward designs that only rely on the semantic meaning of each individual element in hierarchical data and not the content of sibling elements. In this case, however, the sibling directly determines the semantic meaning. We worked with the MDMI group to implement this “linked data” functionality.

Second, MDMI was designed to expect both source and target messages to follow roughly the same hierarchy (e.g., all information on a medication is a child of that medication). C-CDA employs hierarchical information to embed observation details (e.g., medication dose and frequency), whereas PDO is a flat structure that uses a sequence of linked observations (a fact and its modifiers). In order to support modifiers, we worked with the MDMI group to add a feature to ‘flatten’ hierarchical information. Presently this feature is only supported uni-directionally, so it is not possible to move from PDO to other more compact formats using MDMI at present.

Third, the complexity of tracking and testing transformations between three different formats mandated improvements in existing debugging tools and techniques. Using MDMI’s new end-to-end traceability application, we produced traceability reports for these maps. This application takes maps and generates a report on how each data element in the various formats can be traced to the other data elements in the other formats. This tool should be valuable in future MDMI i2b2 projects.

A simplified example of a single medication dispense event in the three formats is shown in Figure 2. The example includes data elements needed to create a valid i2b2 observation as well as information on dispense quantity and instructions. The HL7v2 format is structured around “segments”: the PID segment provides patient information, the RDS segment provides pharmacy dispensing data, the ORC sub-segment provides metadata about the order, and the RXD subsegment provides data on the drug being dispensed. The C-CDA is a deeply nested XML structure that puts all information under a “substance administration supply” event (except the patient ID, which is not shown for compactness). Notice that some pieces of information require even more deeply nested structures under the supply event (e.g., patient instructions are four levels deeper). In contrast, the i2b2 format is a nearly flat structure organized into a sequence of observations. This example shows two linked modifier observations: dispense quantity and instructions. The core components of the observation are repeated for each modifier.

The C-CDA and PDO mappings will be released open source in Open Health Tools with the enhanced MDMI toolkit, which includes the transformation engine and the graphical map editor (http://sourceforge.net/projects/mdmi). [21] The HL7v2 mappings reference proprietary PBM data structures and cannot be released.

Finally, we developed i2b2 ontologies necessary for querying the prescribing, dispensing, and payment data. Because these ontologies are based on the C-CDA, we believe these are the beginning of standards-based medication ontology for i2b2. The medication tree was based on our previously developed RxNorm tree organized by Veteran’s Administration (VA) drug class. [15] Because the dispensing data is in National Drug Code (NDC) format, we used the linkage information present in the Unified Medical Language System (UMLS) to augment our ontology. Specifically, NDC codes for specific drugs are placed as children of the RxNorm drug formulation they are associated with. The ontology for modifiers is a list directly derived from the modifier codes in the PDO. Moreover, the PDO modifier codes are themselves derived from the MDMI Referent Index names. Once the Referent Index is an approved standard, this provides one path toward canonical naming of i2b2 modifiers. Because such a vast amount of medication metadata is included in the dispensing data, for this initial version we prioritized the modifiers most important in obtaining accurate patient medication information: detailed information on the drug, what was dispensed, and refills remaining. Not included are many of the details on credentials, such as NCPDP number (national pharmacy identifier code). More detail on the modifier codes in the ontology is shown in Table 2.

**Validation**

We deployed i2b2 1.7.04 with our i2me2 cell and the RxConnect tools at Boston Children’s Hospital. The patient annotation app was deployed as a web application that can interact with the i2me2 cell. To validate the patient annotation process shown in Figure 1, we loaded synthetic test data in HL7v2 format into i2b2.
We manually augmented the loaded data with prescribing information provided by the i2b2 demo data and used the patient annotation app to view a patient’s combined medications. We also verified our ability to run patient count queries on the imported data in the i2b2 Query Tool.

<table>
<thead>
<tr>
<th>HL7v2 (as XML, with shortened closing tags)</th>
<th>I2b2 PDO (XML)</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;pid&gt;</code></td>
<td><code>&lt;observation&gt;</code></td>
</tr>
<tr>
<td><code>&lt;patientId&gt;8237&lt;/patientId&gt;</code></td>
<td><code>&lt;event_id source=&quot;BCH&quot;&gt;12701&lt;/event_id&gt;</code></td>
</tr>
<tr>
<td><code>&lt;patientIdSourceSystemName&gt;BCH&lt;/patientIdSourceSystemName&gt;</code></td>
<td><code>&lt;concept_cd&gt;NDC:00008084181&lt;/concept_cd&gt;</code></td>
</tr>
<tr>
<td><code>&lt;rds&gt;</code></td>
<td><code>&lt;start_date&gt;2014-02-09T00:00:00.0&lt;/start_date&gt;</code></td>
</tr>
<tr>
<td><code>&lt;orc&gt;</code></td>
<td><code>&lt;patient_id source=&quot;BCH&quot;&gt;8237&lt;/patient_id&gt;</code></td>
</tr>
<tr>
<td><code>&lt;providerId&gt;123456789&lt;/providerId&gt;</code></td>
<td><code>&lt;nval_num&gt;30&lt;/nval_num&gt;</code></td>
</tr>
<tr>
<td><code>&lt;rxd&gt;</code></td>
<td><code>&lt;units_cd&gt;ZZ&lt;/units_cd&gt;</code></td>
</tr>
<tr>
<td><code>&lt;dispenseCodeIdentifier&gt;00008084181&lt;/dispenseCodeIdentifier&gt;</code></td>
<td><code>&lt;modifier_cd&gt;MED:DISPENSEQUANTITY&lt;/modifier_cd&gt;</code></td>
</tr>
<tr>
<td><code>&lt;dispenseCodeNameOfCodingSystem&gt;NDC&lt;/dispenseCodeNameOfCodingSystem&gt;</code></td>
<td><code>&lt;valuetype_cd&gt;N&lt;/valuetype_cd&gt;</code></td>
</tr>
<tr>
<td><code>&lt;dateTime&gt;Feb 8, 2014 12:00:00 AM&lt;/dateTime&gt;</code></td>
<td><code>&lt;observer_cd&gt;123456789&lt;/observer_cd&gt;</code></td>
</tr>
<tr>
<td><code>&lt;actualDispenseAmount&gt;30&lt;/actualDispenseAmount&gt;</code></td>
<td><code>&lt;instance_num&gt;1&lt;/instance_num&gt;</code></td>
</tr>
<tr>
<td><code>&lt;actualDispenseUnitsIdentifier&gt;ZZ&lt;/actualDispenseUnitsIdentifier&gt;</code></td>
<td><code>&lt;observer_cd&gt;123456789&lt;/observer_cd&gt;</code></td>
</tr>
<tr>
<td><code>&lt;prescriptionNumber&gt;12701&lt;/prescriptionNumber&gt;</code></td>
<td><code>&lt;observation_blob&gt;</code></td>
</tr>
<tr>
<td><code>&lt;dispenseNotes&gt;TAKE 1 TABLET EVERY DAY&lt;/dispenseNotes&gt;</code></td>
<td><code>&lt;start_date&gt;2014-02-09T00:00:00.0&lt;/start_date&gt;</code></td>
</tr>
</tbody>
</table>

C-CDA (XML, with some entries on the same line for compactness)

```xml
<section>
  <templateId root="2.16.840.1.113883.10.20.22.2.1"/>
  <entry/><substanceAdministration><entryRelationship>
    <supply>
      <templateId root="2.16.840.1.113883.10.20.22.4.18"/>
      <id assigningAuthorityName="BCH" root="12701"/>
      <effectiveTime>20140208240000</effectiveTime>
      <quantity>
        <value>30</value>
        <unit>ZZ</unit>
      </quantity>
      <code>
        <code>00008084181</code>
        <codeSystemName>NDC</codeSystemName>
      </code>
      <author><assignedAuthor><assignedPerson>
        <id assigningAuthorityName="DEA" root="123456789"/>
        </assignedPerson></assignedAuthor></author>
      <entryRelationship><observation><text>TAKE 1 TABLET EVERY DAY</text>
      </observation></entryRelationship>
    </supply>
  </entryRelationship></substanceAdministration></entry>
</section>
```

**Figure 2:** An example of a medication dispensing event in the three data formats: HL7v2, C-CDA, and i2b2 PDO. HL7v2 is presented in a condensed XML format for readability – the actual format is a very terse pipe-delimited structure that is not human-readable. Notice the record-based format for HL7v2, the deeply-nested document-based format for C-CDA, and the denormalized PDO format (wherein a single logical observation requires multiple i2b2 “observations”).
Table 1: Mapped elements that translate to i2b2 modifiers and their HL7v2 location and i2b2 modifier code. The actual values are stored in appropriate columns in the i2b2 observation fact table. Shown here are codes for a dispensing event.

<table>
<thead>
<tr>
<th>Name</th>
<th>HL7v2</th>
<th>PDO modifier code</th>
</tr>
</thead>
<tbody>
<tr>
<td># of days filled</td>
<td>ORC/Duration</td>
<td>MED:DISPENSEDURATION</td>
</tr>
<tr>
<td>Pharmacy Name</td>
<td>RXD/DispenseProviderFamilyName</td>
<td>MED:DISPENSEORG</td>
</tr>
<tr>
<td># of refills remaining</td>
<td>RXD/NumberOfRefillsRemaining</td>
<td>MED:REFILLSREM</td>
</tr>
<tr>
<td>The type of pharmacy</td>
<td>RXD/DispenseProviderIdentifierTypeCode</td>
<td>MED:DISPENSETYPE</td>
</tr>
<tr>
<td>Actual # of pills dispensed</td>
<td>RXD/ActualDispenseAmount</td>
<td>MED:QUANTITY</td>
</tr>
<tr>
<td>Patient Instructions provided on label (the &quot;sig&quot;)</td>
<td>RXD/DispenseNotes</td>
<td>MED:INSTRUCTIONS</td>
</tr>
</tbody>
</table>

Discussion

This represents a successful validation of an approach to integrate important non-traditional data sources to enhance medication lists for post-market drug surveillance. In this work, we described the successful end-to-end tests of a toolset that combines PBM data with i2b2 data, which is presented to patients for corrections via tablet-based patient annotation. This study also describes how complex medication information can be represented in three very different data formats.

The project team is presently recruiting patients for studies that examine the data quality of the medication list using this toolset. The current goal is to identify discordance between the clinical medication list and the patient-generated list (the output of the annotation app). Then, techniques will be developed to assess adherence from the discordance.

This is a generalizable approach to importing a variety of medication data into i2b2 via C-CDA messages. Our inclusion of PBM data enables a novel data source for patient-performed medication list annotation. Also, the enhanced MDMI toolkit and the maps developed for this project offer a starting point for data interoperability between i2b2 and a variety of standardized or custom data formats. The MDMI toolkit and maps for i2b2 and C-CDA will be released open-source shortly on the Open Health Tools website, and non-PBM-specific components of the i2me2 codebase will be released open source by the completion of the project.

Limitations and Future Directions

The patient-facing app functionality for loading reconciled medication data into i2b2 remains under active development. A standard ontology for patient reported medication information is being developed. The initial version of the app app uses i2b2 PDO format, but the final version will use the Fast Healthcare Interoperability Resources (FHIR) format, which is considerably more compact as well as more efficient to transmit and parse. Upon completion of full end-to-end testing using synthetic data, we will proceed to test our approaches using real-time data from consented subjects.

Perhaps the greatest technical limitation of this toolset at present is the reliance on PDO format for client-facing transactions. While PDO format is flexible, the inherent denormalization creates highly verbose output. Therefore it is not an efficient transmission format between the app and i2b2. In the future, we will be leveraging previous work in implementing SMART on i2b2, including addition of FHIR-based SMART messaging.

Conclusion

We have demonstrated the ability to incorporate a novel data source (pharmacy dispensing data) into the medication list of the i2b2 EHR data warehouse. The C-CDA approach is generalizable and uses standard data transmission format. This combined list is then presented to patients for annotation, adding a second novel data source (the patient). These tools will be released open source and they are currently being used at Boston Children’s Hospital to
power studies on the concordance of medication lists across data sources. This work enables important research to improve medication list quality for drug surveillance and to improve patient drug adherence.

Acknowledgements

Thanks go to: Lori Phillips on the i2b2 team for contributing her expertise in ontology design and development; to Florence Bourgeois in the Children’s Hospital Informatics Program for her leadership of study design and research applications of the i2me2 platform; and, to the MDMI team (Ken Lord, Sean Muir, Sally Conway, and Gabriel Oancea) who worked closely with us to make sure all needed functionality was available to us. This work was supported by NIGMS/NIH grant R01 GM104303.

References


Long-Term Engagement with Health-Management Technology:
 a Dynamic Process in Diabetes

Predrag Klasnja, PhD¹, Logan Kendall, MHA², Wanda Pratt, PhD³,
Katherine Blondon, MD, PhD³
¹University of Michigan, Ann Arbor, MI; ²University of Washington, Seattle, WA;
³University Hospitals of Geneva, Switzerland

Abstract

Diabetes management is a complex, dynamic process that is largely incumbent on patient choices and behavior. We explore how health-management needs—and the needs for technological support—change over time for individuals with diabetes. Through interviews and a focus group, we found that after initial diagnosis, individuals face acute information needs and chiefly turn to mobile applications and Internet resources to help understand the diabetes-specific factors that affect their health. Over time their focus shifts from highly regimented routines to more flexible ones that enable them to maintain a quality of life. Our results suggest that long-term engagement with health technology does not necessarily require continuous, sustained use: routine disease management could lead to a decrease in use, until a new event occurs. Our findings point to a need for tools that help patients with diabetes to effectively manage their health as their bodies, treatment and circumstances change over time.

1. Introduction

Diabetes is a common, chronic disease that affects 346 million people worldwide, and is predicted to become the seventh leading cause of death by 2030¹. There are two main types of diabetes: Individuals with type 1 diabetes (T1D) have a deficit in insulin, whereas those with type 2 diabetes (T2D) have an increased resistance to insulin. Diabetes leads to increased risk of cardiovascular events, is the leading cause of kidney failure and amputations, and is a major cause of blindness. Patients that can reduce their average blood glucose level (HbA1c), control their blood pressure, and treat their cholesterol can delay the onset of these complications. Yet only one in eight individuals achieves these three goals². Such low rates of adequate self-management suggest that the standard model of care delivery consisting of short, infrequent visits with care providers conveys insufficient support for effective health management, including managing medications, improving diet, controlling weight, increasing physical activity, and smoking cessation³.

In recent years, mobile technology has emerged as a promising way to provide additional support for self-management⁴. Adoption of smartphones and other mobile devices has grown exponentially in all age groups and socioeconomic classes⁵ and applications for diabetes and wellness have flourished⁶. The available smartphone applications for diabetes are diverse, and allow individuals to track activities, medications, or diet, as well as to visualize the results in graphs. These types of applications offer new ways of supporting the challenge of diabetes self-management, and they have shown promise in clinical trials. A meta-analysis of 22 trials assessing the effect of mobile phone interventions on glycemic control showed a reduction of HbA1c of 0.5% over a median of 6 months’ follow-up duration⁷. This effect was shown to be greater in individuals with T2D than in T1D⁸. Early work on the use of mobile phones for diabetes management studied the use of phone calls or text messaging⁹. Recent studies focused on tracking applications (insulin, diet, blood glucose, and weight, for example), educational content, or integration with other systems, such as personal health records and social media⁶. Ralston et al found that the use of secure messaging with providers improved HbA1c results by 0.7%¹⁰. Combining mobile device and provider support, a randomized controlled trial with WellDoc Diabetes Manager™, a FDA-approved application that offers automated clinical coaching based on the data that patients track and supports sharing of that information with care providers, showed a 1.2% decrease in HbA1c with web- and mobile-based tools compared to usual care over a year⁶. This result suggests that mobile technology might be even more effective as a part of a comprehensive suite of technological tools.

Whether these preliminary efficacy results are an accurate picture of the benefits that mobile technology could have on diabetes self-management is unclear. In studies with a longitudinal design to assess the clinical impact of technological interventions for diabetes self-management⁷, researchers often classify irregular use of technology as a failure in use. However, the lack of use could be due to a number of factors, including a mismatch between the evaluated technology and the current stage of the patient’s condition or technology design that doesn’t enable patients to sufficiently adapt the application to their specific needs. For instance, Chen has found that in spite
of a similar set of self-management needs in newly diagnosed individuals, responses to these needs vary depending on the individual’s physiological, social, and personal circumstances. If the application that is being evaluated does not adequately match patients’ current self-management needs, patients will use the system irregularly or stop using it altogether. In such cases, trial results that report average effects for the whole sample would show lower efficacy than would be found in a subsample of patients whose self-management needs were well matched by the evaluated application. Given that the large majority of current mHealth tools for diabetes rely on a single self-management strategy, tracking, the possibility of a mismatch between patient needs and available support cannot be ignored.

While research in health sciences has focused on assessing efficacy of self-management tools for diabetes, work in human-computer interaction (HCI) is exploring new approaches for technological support for behavior change, such as personalization and sense-making. For example, Mamykina et al. have focused on the development of problem-solving skills for diabetes self-management. Drawing on the construct of sense-making, they developed a mobile-phone application, MAHI, which enables patients to collect contextual information related to their glucose measurements (e.g., what they ate, where they are, etc.). The information is uploaded to a secure website where patients can reflect on patterns in their information and discuss their data with a diabetes educator. Frost and colleagues took a similar approach: to facilitate reflection on patterns in glucose measures, the researchers used photographs to annotate glucose readings to enable patients and their providers to better understand each individual’s values.

The work in HCI and in health sciences has only begun to explore how patients’ needs for self-management technologies evolve over the course of the disease and how we can design technologies to account for the evolving nature of diabetes self-management. For example, Chen has suggested that seeking of health information is cyclical rather than linear. If the needs for self-management support substantially change over time, the benefits we are seeing from self-management tools might be less than they could be, in part due to the potential mismatch between the type of technology being tested and the specific needs of the patients participating in the study. To avoid such mismatches, we need to better comprehend individuals’ evolving self-management needs and the ways that technology can support them over time.

In this paper, we begin to close this gap in understanding how diabetes patients’ health-management needs change over the course of the disease and how they use various technologies and devices to support those changing needs, using a grounded theory approach. Our findings can inform the design of technologies that account for the shifts in self-management priorities brought about through the progression of the disease, changes in treatment, and the patients’ own evolving skills and understanding of diabetes self-care.

2. Methods

2.1. Data collection

After receiving approval from the Institutional Review Board, we recruited adults with T1D or T2D through flyers in Diabetes Clinic of the University of Washington, convenience sampling, and a diabetes support group from September 2011 to October 2012. We excluded individuals with gestational diabetes because of its limited duration, as well as those on dialysis, because they see their care-providers much more frequently than other individuals. Participants received $25 for participation. We collected participants’ socio-demographic characteristics, and information about their technology and healthcare use. We chose not to perform purposive sampling by disease duration, as the relationship between disease duration and disease stage is not clearly established.

We led in-depth, semi-structured, individual interviews with 11 participants. All sessions were audio-recorded and transcribed for subsequent analysis. The sessions ranged from 35-100 minutes. We asked individuals to describe their current diabetes management, their trajectory since they were diagnosed, and their use of technologies, such as their use of mobile devices and the Internet (e.g., forums, blogs, other websites). We also explored their perceived role in coordinating the care they receive. Finally, we asked them about their perceptions of barriers and motivators for diabetes self-management. We also conducted a two-hour focus group with four participants, three of whom had been interviewed individually: three individuals with T2D (diagnosed in the past two years) and one individual with T1D (diagnosed 20 years ago). During this session, participants discussed their disease, their evolving needs, barriers and motivations, and discussed their use and need for supportive technologies for diabetes.

2.2. Data analysis

Two coders each analyzed the transcripts of the individual interviews and the focus group in Atlas.ti 7 software, using open coding to establish prominent themes encountered in the data: the evolution over time of the disease and
its management; the tracked information; the comprehension, availability and usefulness of technologies such as glucose meters and their companion websites, insulin pumps, blood pressure monitors, smartphones and tablets, online forums and social networking. Codes were developed for the disease trajectory, modifications in symptoms, treatments, barriers and motivations for self-care. The two coders iteratively compared their coding schemes, revising or merging them as they discovered common and contrasting themes. We took into consideration duration and type of disease for this analysis. Reaching saturation allowed us to terminate the data collection.

3. Results

3.1 Participant characteristics

Across the interviews and focus group session, we enrolled 12 diabetes participants (six T1D and six T2D patients), who had been diagnosed for periods ranging from 6 months to over 20 years. There were five male participants, four of whom had T2D. One patient with T2D used insulin. The patient characteristics are presented in Table 1. The age range in our sample is representative of T1D and T2D demographic characteristics in the general population. Three T2D participants were Asian, one T2D participant was Black, all other participants were Caucasian.

![Table 1. Participant demographics and disease characteristics.](image)

<table>
<thead>
<tr>
<th></th>
<th>Type 1</th>
<th>Type 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>6</td>
<td>6*</td>
</tr>
<tr>
<td>Female: male</td>
<td>5:1</td>
<td>2:4</td>
</tr>
<tr>
<td>Age range</td>
<td>18-34 yrs</td>
<td>35-64 yrs</td>
</tr>
<tr>
<td>Duration of disease by participant</td>
<td>27 yrs (P3), 12 yrs (P6),</td>
<td>17 yrs (P1), 0.5 yrs (P9),</td>
</tr>
<tr>
<td></td>
<td>19 yrs (P4), 23 yrs (P8),</td>
<td>21 yrs (P2), 1.5 yrs (P11),</td>
</tr>
<tr>
<td></td>
<td>8 years (P5)</td>
<td>1.3 yrs (P10)</td>
</tr>
</tbody>
</table>

*Only one individual with T2D using insulin

3.2. Evolution in health management needs

A key finding from our work is that patients’ priorities and needs for self-management support change substantially over the course of the disease. In this section, we describe patients’ needs and priorities in the following phases: (1) the learning phase in the initial period after diagnosis, (2) the stabilization phase, when they gained gain confidence, (3) the relearning phase, when they need to learn how to respond to changes brought about by the disease progression and adjustments in treatment, and (4) the expertise phase revealed during long-term management.

3.2.1. Learning phase: Developing understanding after diagnosis

Our findings indicate that the initial period after diagnosis is characterized by acute psychological distress, a need to learn about the disease, and to adopt new routines and health behaviors to control glucose levels. During this period, patients make heavy use of tracking tools and of educational websites, but find the available information insufficient.

Participants find this initial stage difficult, both at a physical and psychological level. Two participants consulted mental health professionals to help cope with their new chronic disease. Diabetes has a negative reputation, with its ban on sweets, its amputations, the painful pricks and injections and medications. Such perceptions are often buttressed by friends or family members with diabetes who have presented these complications. For one participant (P1, T2D) the moment of diagnosis was “the branding, the shock” and her initial thought was “I’m a failure.” The DAWN study found that 85% of individuals reported a high level of stress (shock, guilt, anger, anxiety, depression, and helplessness) after receiving this diagnosis. The difficulty of finding practical information heightened the feeling of being overwhelmed by the disease in the early stages as they tried to understand the expectations for their new lifestyle: “I wanted something early on, I just want to know if I can eat five crackers, if that’s all I have to eat, I’ll eat crackers for 90 days. [...] Give me a recipe, I can follow it and not have to think about it so much.” – FG, P7 (T2D)

In addition, a need for intensive learning characterizes this initial period as participants search for answers about diabetes and discover how to incorporate management strategies to get better and prevent complications. Self-management at this stage focuses on understanding the basics of the disease, and changing daily activities, such as nutrition. Lifestyle changes are particularly important in type 2 diabetes, where portion control, healthy food choices, and physical activity are initial steps to decrease and stabilize glucose levels. As FG, P7 (T2D) noted,
“Going back, early on, it’s food. You know, you have to figure out what you can eat, and then it’s exercise and then it’s the combination of everything and then it’s working it all in the lifestyle stuff too.”

Although some information seemed difficult to find on the Internet, our subjects found comfort in its constant availability as a resource: “It’s not just ‘go to this class that’s taught in two weeks’ but ‘here’s where you could start online tonight when you’re freaking out at two o’clock in the morning’” – Intv, P5 (T1D). However, they also described how time-consuming and impractical websites were to use, particularly at early stages. This was particularly the case for a nutritional website which provides detailed breakdowns of nutrient data for an extensive selection of foods; participants described how it was difficult to utilize this data when combining several foods together. “There’s not a website that says ‘eat this - you can eat this, in this proportion’, none of that. And the same thing with exercise. You know, they say ‘exercise’ but then they don’t give you – they don’t just say ‘10 minutes per meal’.” – FG, P7 (T2D). The challenge of controlling glucose levels with food and exercise is compounded by the lack of concrete, actionable plans that new patients can adopt to get them started. Although many websites, handouts and textbooks about diabetes exist, patients expressed frustration with finding concrete information to guide daily choices. Participants found the available food plans bland, describing the need to adjust to, “taste, lifestyle, you know, even down to the city, because [most] is tailored to the Caucasian diet right now.” – FG, P9 (T2D)

At this early stage, patients describe strategies to help adopt new routines for disease management. They integrate new behaviors like tracking, glucose testing, and they limit their food choices to those that were successful in the past: “I was tracking down to individual ingredients I was putting into my recipes and just breaking it down to what I can. You know, we were measuring every two hours after a meal, doing exactly what they were telling us to do, wake up: check, after breakfast: check, two hours before lunch: check, two hours after lunch: check. So it was a whole routine for about the first month, two months.” – FG, P7 (T2D)

Detailed tracking helped participants learn how different foods and routines had an impact on their blood sugar levels. However, in this early phase, participants would also simplify their lifestyle and choices to activities that they already knew how they impacted their bodies. One participant described always going to the same fast food restaurant while on the road, “because I know exactly what that’s going to do to my blood sugar” – Intv, P6 (T1D). Restricting options is difficult, but it saves time and decreases anxiety about glucose management.

Individuals use intense tracking to learn how to interpret many numbers, including blood glucose results and carb counts and the effect of behaviors on their body. Patients might also start new medications and want to observe the effects of those drugs. Some participants used pen and paper or Excel spreadsheets, while others turned to their smartphones for tools to assist them, with options for averages and other analyses. They particularly appreciated the ease of data retrieval and seeing associations between trends in carbs and glucose levels: “The nice thing about having it on the phone is I can just give [my providers] my last 90 days, without having to scroll back through a bunch of spreadsheets and paperwork and calculate the number. I can just quickly at a glance say where I am” – FG, P7 (T2D)

Even though some of our participants developed tracking strategies on their own, they wished for more guidance from their care-providers for electronic resources (websites and smartphone apps). A number of websites and apps are available for diabetes, nutrition, physical activities and weight, yet providers do not seem familiar with these resources. Patients expect their providers to suggest and advise them about these resources to ensure that patients are accessing high quality information: “I feel like the good [providers] are also learning to go online or at least realize that there are conversations online [...] knowing good resources whether it’s diabetes.org or the TuDiabetes.com network. So giving patients a starting place and not making them go to Google.” – Intv, P5 (T1D).

Supplementing their interactions with providers, our participants also leveraged peer support from other individuals with diabetes, particularly at this early stage of disease. The experiences of peers can provide information not given or emphasized by care professionals: “My friend actually was the one who told me to take my insulin earlier to offset like spikes and stuff. [...] She was the only person who said that to me, I guess. She’s kind of clued me in, that’s helped a lot.” – Intv, P10 (T1D)

During this initial phase, then, patients face the challenge to understand how their bodies respond to food, medication, and exercise, and, based on that knowledge, to come up with daily routines that help stabilize their blood sugar levels. Tracking tools are a key support for this process, as are online resources such as blogs and forums where patients can get advice and tips from other patients.

3.2.2. Stabilization phase: Building confidence

As individuals make efforts to adjust to new lifestyle choices, they progressively build skills in and confidence about diabetes self-management. They realize that they can delay the onset of complications and that their actions can bear
positive results, reflected in their HbA1c numbers or feedback from their care-provider. Patients also master the diabetes lingo, as P8 (T1D) described: “Breakfast was one protein, two starches, one fat, one dairy, one fruit...” After they gain initial control, patients also begin to slowly integrate new behaviors into their daily routines:

*I think that initial first three months was just getting an understanding of “can we get these numbers down,” “what is it going to take” and then now that we got them down, it seems like it’s okay to go out and have a couple of drinks with your co-workers after work or happy hour, and do stuff. [That shift occurred] after the first A1c that went down.* – PG, P7 (T2D)

During this stabilization phase, individuals gradually lose interest in tracking their diet or other activities, because they have adopted routines that generate satisfactory results. They have discovered which foods to eat in what proportion, and have developed skills at interpreting glucose results and at estimating carbs in common foods.

[I made] a master list of all the items that I could possibly eat and the glycemic load and the calories [...] For the first 60 days, [our glucose results] were always low because we knew exactly what we ate. So then after a while it was like it was the same thing we’re eating every day, you know what it’s going to be. So after that I just stopped [tracking]. – FG, P7 (T2D)

*I can usually look at things and [...] probably 70 - 75 % of the time I get at least in the ballpark.* – Intv, P8 (T1D)

Individuals gain confidence in their skills to assess carbs, or predict their response to certain activities. They might even search for ways to fine-tune their glucose results through detection of complex interactions among emotions or physical activity. They recognize that mobile technologies can provide a lot of information, but do not use them as much at this point.

*Some apps will tell you like if say, you know, three strawberries, or whatever, they will tell you how many carbs that is. But I kind of just eyeball it and you know, from experience, just sort of trial and error with different things* – Intv, P4 (T1D)

*There have been a couple of times where I come out of the pool and been really low – it was at the beginning when I didn’t know exactly how it would affect me but I feel like, generally, at least now, I’m able to kind of predict what’s going to happen and I take glucose tablets with me when I run and bike and that kind of stuff.* – Intv, P10 (T1D)

The key feature of the stabilization stage, then, is that patients have managed to internalize knowledge that previously required them to use external tools. By better understanding their bodies and the ways in which factors that affect their sugar levels interact, they are able to “eyeball” different types of foods and predict the effects that eating something will have on them. Consequently, intensive tracking technology became less useful at this point.

### 3.2.3. Relearning phase: Responding to changes

Some participants described feelings of guilt when their glucose results were less satisfactory. Consideration about staying healthy for their loved ones only compounds this feeling: “I think I just start feeling guilty because [...] I was not managing my diabetes the way that I should. [...] And it’s unfair to my husband to be so nonchalant about it. So there’s that guilt that kicks you into action in a positive way.” – Intv, P6 (T1D)

Glucose levels can start reacting differently over time for various reasons. Even though individuals have learned how life choices affect glucose results, how various factors affect a person’s body can change over time. Our participants described a continuous need to re-evaluate their lifestyle in order to understand changes in glucose patterns.

Some changes in glycemic control are necessitated by the changes in the body itself, often as a result of the disease progression: “It’s [been] two years now that I’ve had diabetes; it’s changing pretty significantly now. And so I’m having to like re-learn things that I thought I kind of at least had a feel for the last couple of years. [...] But I’ve noticed that I need more insulin for any of the carbs that I do eat. I need to take them earlier, like 15 minutes before. [...] Generally speaking, I need more insulin, so any of the numbers I did learn for the foods that I typically eat, I’m kind of relearning now, how it affects me now” – Intv, P10 (T1D)

The need for relearn glucose control can also be brought about by changes in treatment such as new treatment regimens or a new device, such as a continuous glucose monitor (CGM) or insulin pump, which can drastically alter diabetes self-management. “Before I got my CGM, it was really a guessing game. It was very much like, you think that you know what’s going on and what’s happening. [With a CGM,] you have this meal every single morning and this is what your blood sugar did every single morning. What was the difference with this morning versus others,
you know, because everything else is the same?" – Intv, P4 (T1D). Finally, relearning can be necessitated by changes in the patients’ lives, such as travels or a new job: “To try to eat on a regular schedule, to try to eat foods that are appropriate on a regular schedule are not simple on the road” – Intv, P2 (T2D).

Whenever patients notice that their bodies are responding in unexpected ways, they must determine whether the change in glucose response is a one-time event or a pattern that would require an adjustment of diet or insulin regimen. Individuals adopt different strategies to answer that question: some will look at previous days, others rely on their memory, and yet others will again start tracking their food, glucose, insulin, and activities more closely: “I don’t really track much unless I am specifically looking for a pattern.” – Intv, P4 (T1D)

To monitor changes in glucose patterns, providers may ask individuals to track their diet and glucose in preparation for visits. This helps assess if there is any gradual change in trends, and provides concrete points for feedback if abnormal results are found. Although care-providers can provide insight into some abnormal results, they might not have the contextual and lifestyle information needed to interpret all results. Tracked results without context can lead to misunderstandings and frustration for the patients, while limiting the provider’s ability to provide feedback... As P4 noted, “The chart of all my blood sugars, it’s, like, each number has a story behind it. And it’s like, I remember that meal or I was sick or, you know, there’s so much more than just numbers that goes into it.” – Intv, P4 (T1D)

Learning to manage diabetes is not a one-time shot, but rather a continuous process. As the circumstances change, patients have to relearn how their bodies react to different factors that were once a known quantity. The need for external support often goes up in these situations. Patients will turn to their clinicians as well as technology to regain control of their glucose levels.

3.2.4. Expertise phase: Managing for the long-term

Long-term management of diabetes is important, because effective diabetes management can reduce complications. Yet, the disease does not remain patients’ main focus and instead they begin to prioritize quality of life. Over time, patients develop a nuanced and detailed understanding of their illness that leads to balancing glucose management with the desire to explore new foods and activities.

This quality of life: like I know that I can spend so much time and be so diligent and write everything down and do what you were doing in the beginning. But is that really worth it in the long run? [...] I test on average 10 times a day. [...] But that’s seven seconds out of my life, the tracking. So that’s definitely a goal of mine, [...] finding that balance a little bit better and maybe doing a little less work, but still getting the good results – FG, P4 (T1D)

I mean, you’re not going to live your whole life watching, counting calories, every grape that you eat. I mean I’m getting really close to that point now where [...] I’m not thinking about my meals every day. I’m content with like, if I miss a snack so, I’ll make it up at dinner – FG, P7 (T2D)

As individuals grow to live with diabetes, there is a point when they understand enough about what behaviors they should adhere to, but also realize that they are not enjoying life as much as they would like to.

Individuals discover over time that the metabolism of glucose can be intricate and sensitive to a range of factors: “I’ve had [diabetes for] almost 20 years now, it’s crazy. But, I’m still learning everyday just what I can and cannot eat and – because even if you go straight off carbs, fiber, glycemic index, like there’s other things that go into it. [...] And glycemic load is a little bit dependent on the meal. So like if you eat like grapes by themselves, they might have a certain glycemic load but if you eat them with cheese, it’ll change it.” – FG, P4 (T1D). In addition, Type 2 patients not on insulin therapy frequently taper their glucose monitoring to avoid the unpleasant finger prick required. Although this can be patient-initiated, providers also encourage this when results are satisfactory: “Since I’ve been a diabetic for over 20 years, I just do it in the morning. I don’t watch it other times of the day that usually ends up being a good reflection about how things are being handled.” – Intv, P1 (T2D)

As individuals loosen their dietary restrictions, they may wonder how this affects their glucose and initiate a short period of glucose tracking to see if they are still on track: “I check [my blood glucose] once in a while when I do try something different, when I eat something different or, you know, if I’m feeling awkward or something.” – Intv, P7 (T2D). While some individuals used electronic spreadsheets or handwritten notes, others used smartphone weight-tracking applications that provided detailed information about the breakdown of foods.

3.3. Social and emotional impacts across phases

Diabetes, like many chronic conditions, can lead to a variety of different emotions. The pressure of handling food choices, calculating carbs and insulin doses, and testing glucose on a daily basis can lead to anxiety in some
individuals: “I want to be able to eat food. Like I want to able to try food and not be really scared of everything I put in my mouth” – FG, P4 (T1D). These emotions also occur in T2D individuals, even without insulin therapy: “There’s really a mental pressure. You’re talking about the constant thing. It’s a constant mental pressure on me to always keep track of what I eat, how much I eat, when I eat, and on and on” – FG, P11 (T2D)

For many, the motivation to keep up with the many daily requirements of one’s regimen is to stay healthy and delay the onset of complications or need for new medications. Many individuals are influenced by a parent or relative who has experienced, or even succumbed to these complications. Fear of consequences can drive motivation for self-management: “The negative aspect would be knowing that eventually, [diabetes] will catch up. It’s a non-stopping disease and eventually it will catch up with me and I don’t know why, I don’t know what will happen to my bones or my eyes or skin or heart or something will eventually go. […] And that knowing is sometimes hard to handle, that knowing. So my own father passed away because of complications of diabetes.” – FG, P11 (T2D)

The balance between highly controlled self-management and spontaneous activities is a fine line. Individuals can push themselves by setting high standards with the risk of being too demanding, or they can feel pressure from their care-providers to improve numbers. Performance in disease management can be a sensitive topic, as P8 (T1D) describes: “While I’m willing – able and willing to talk about my diabetes, I’m not – I don’t want to have any perception of being judged.” This may be why some individuals feel a particular affinity with peers, particularly care-providers who also have diabetes. There is a shared understanding that behaviors cannot be perfect all the time and less than ideal results do not invite critical judgment.

This doctor, who is a type 1 diabetic that I met at diabetes camp, […] she was the only doctor I’ve ever met that like I felt like really understood and really just made it okay to eat ice cream if you knew what to do and… It’s okay if you have high blood sugar every once in a while, just keep on going, and keep figuring out. – FG, P4 (T1D)

To an outsider, diabetes appears to be a burdensome disease with many requirements for testing, medications, lifestyle changes and restrictions in diet. Yet our participants had a very positive approach and expressed an acceptance with having to live with this condition. Individuals with T1D tended to describe the disease as the way their body was, whereas T2D participants considered it as a wake-up call because they had become so unhealthy.

I still have diabetes, and I still have to do this every day. […] That’s the way I am as a person – Intv, P4 (T1D)

I always tell the doctor that being diagnosed with diabetes was the best thing that ever happened to me, because it stopped me in my tracks. I wish I did that I’d been diagnosed twenty years ago because I would not be in the same health situation as I am now. – Intv, P12 (T2D).

4. Discussion and design implications

Our finding that self-management needs change over time aligns well with the Corbin and Strauss’s Chronic Illness Trajectory and suggests that technology will play different roles at different stages of the disease. In particular, based on our studies, we identified four functions where technology could play a supportive role: understanding the new disease, responding to changes in times of stability, improving communication, and tailoring to individual motivations and needs.

4.1. Understanding the new disease

Supportive technologies can play a particularly useful role at the onset of disease, as patients create their mental model of the disease and its management. Mobile technologies can guide initial lifestyle choices of nutrition, physical activity, glucose monitoring or medication in three ways: (1) Tools can facilitate tracking by guiding patients to easily capture a full range of factors that affect glucose levels. By taking advantage of sensing and lightweight self-report, a new generation of tracking apps could enable patients to log, with little burden, not only glucose levels and food intake, but also sleep, stress, physical activity, and other potentially relevant influences on glucose variations. (2) Applications can support learning by enabling patients to understand the interactions among the factors they are tracking. This data interpretation could be achieved through the use of visualizations, coaching (e.g.,), games, and through the use of machine learning algorithms for automatic pattern detection. For instance, a tracking application could not only visualize patterns in historical data, but also use logged data to provide visualizations of likely glucose changes over the next couple of hours—visualizations that could help the user understand and make more informed choices about different foods, activities, or types of insulin. Such visualizations could help the user to get an intuitive understanding of, for example, why foods with the same amount of carbohydrates can affect their bodies so differently or on the duration and effect of the different types of insulin. (3)
Finally, mobile applications can support skill development through the use of video clips, for instance, to teach patients how to perform a foot exam.

4.2. Responding to changes in times of stability

Participants indicated that intense tracking and diet restriction occurred mainly in the early stages of disease. As individuals grasp how to make healthy choices and adopt healthier behaviors, the perceived benefits of intensive tracking are surpassed by the cost in time and effort. While ongoing glucose tracking is recommended for insulin use, it might not be necessary for those who do not need insulin. In addition, as patients learn to estimate content of different types of foods, regular carb tracking becomes less useful. At this stage, the burden of tracking could be decreased for insulin users by switching to a lower-intensity mode for routine tracking, where information is added to the patient’s log automatically (via the glucose meter and sensors), while maintaining enough data to enable the system to detect significant pattern changes that need to be brought to the patient’s attention.

As the disease stabilizes and the concerns about quality of life take front stage again, individuals begin to break out of strict routines and experiment with new foods and activities. One role supportive technology can play at this stage is to support such experimentation while helping individuals to maintain good glucose control. One way to do this is through the creation of a personal knowledge base that patients can use to track their glucose responses and insulin use when they try new foods or activities. Such a system could help improve future decisions through rapid retrieval of prior personal experiences and may help reduce the anxiety of decision-making under uncertainty (P4, T1D and P11, T2D). This is also an area where a large-scale, patient-reported repository could be particularly useful. If a system knew insulin sensitivity parameters for different users, it could automatically use one user’s data to recommend insulin dosages for the same food or activity to other users. Practical applications of this crowdsourcing framework include insulin recommendations for areas with inadequate standardized information available such as restaurant menus, athletic activities, travel, and other activities that affect patients’ glucose levels.

Finally, one challenge in the period of stability is how to detect and communicate abnormal test results. Not all abnormal results require changes in behavior. Machine learning could help detect repeated anomalous events over a short period of time and prompt the user to make an appropriate change, such as changing the basal insulin rate. In our study, we found that the way to communicate a need for change is important. Warning alerts in response to a missing or abnormal glucose measurement may aggravate an individual’s feelings of guilt or may be perceived as a judgment. Future research could explore the use of more subtle cues such as changing the color of the phone’s wallpaper to provide a gentle indication to examine potentially concerning changes in glucose values.

4.3. Improving communication

When individuals have a better understanding of diabetes and its management, they can select relevant information to report for in-depth discussions and useful feedback with their providers. P4 explains how there is a “story behind each [glucose] number”, a lens through which to interpret abnormal results. Self-management technologies should allow individuals to easily capture contextual information that might be relevant for interpreting their glucose results. Smith et al.’s work on photo annotation of glucose data is a step in the right direction. Easily connecting information about location, recent sleep history, workload, stress levels, etc. with specific glucose readings would further support learning and sense-making that can occur in discussions with healthcare providers.

Our findings showed a gap between patient expectations and provider guidance for websites and apps related to diabetes. Participants’ providers welcomed various formats (paper, email or apps) for patient-reported data and appreciated patients’ technological expertise, but they were not a good resource of technical information themselves. Creating a well-maintained database of diabetes-related resources and their reliability could improve patient care.

Sharing with and learning from peers is important. Individuals with a disease become experts in that disease. Our participants readily turned to peers for practical information, such as food recipes or management of insulin with a new type of sport. Peers bring in a different type of information than the providers, because they “get it”, and grasp the challenge of having to always be on top of the disease management. Future technologies could make peer exchange of actionable information and social support easier by integrating a social component with tracking tools and enabling patients to seek support (e.g., on understanding why their glucose responded in a certain way to a particular food or activity) in the context of the information that prompted the need for support. How such social functionality can be designed in a privacy-sensitive and unobtrusive way is an open research challenge.
4.4. Tailoring to individual motivations and needs

Motivations for diabetes self-management vary: for some people delaying the onset of complications is a motivator, whereas for others that same thought creates anxiety. Likewise, for some, the mental burden of disease is increased with tedious data entry and tracking, or bothersome alerts, whereas for others, technology simplifies data retrieval and reminders serve their purpose, allowing patients to concentrate on other activities. Such differences can lead one patient to abandon the same technology that another patient finds invaluable. For these reasons, technologies intended for long-term not only need to support simple customizations—enabling or disabling of reminders, for example—but may need personalization at a much more fundamental level. An application that had a robust user profile that included user’s goals, motivational orientation (e.g., whether the user is promotion- or prevention-focused), attitudes toward illness, and other similar factors, could adapt its behavior to the user’s need at a deep level—employing motivational strategies to which the user would be most receptive, framing glucose results in terms that would not be discouraging, and using personalized content (e.g., user’s professional goals) to support health behaviors. How to construct such rich profiles with minimal user burden is an important research question.

Just as patients’ motivations vary, so do the situations in which they have to perform health activities. Such situations—locations, people patients are with, etc.—can act both as facilitators of and barriers to effective health management. Technology can further support patients’ self-management by helping them create implementation intentions—plans when, where and how they will perform health-promoting activities or resist activities that hurt their health. Given enough data from sensors and the information that the patients themselves log, technologies could both help patients to discover situations that influence their health-related activities and incorporate those situations into effective implementation intentions that can strengthen their health practices.

By taking into account how self-care needs change over time and the individual differences that shape diabetes management, future technologies could help patients with diabetes to manage their disease effectively over the long-term, contributing both to their health and quality of life.

4.5 Limitations

Although our participants are representative of the two types of diabetes, a limitation of our study is the small size of our sample. Due to disease type-related and individual differences in approaches to diabetes self-management, the full range of needs for supportive technologies may not be entirely covered in our sample. Self-selection to participate in studies can be a bias, particularly as most of our participants had outstanding diabetes management.

5. Conclusion

With several short-term studies showing promising results for mobile technologies in chronic disease management and in particular diabetes, we need to explore how to establish long-term engagement with these technologies. Where self-management evolves over time alongside disease trajectories, engagement with technology also becomes a dynamic process. Our findings suggest that the design of tools for diabetes that support long-term engagement should allow periods where the individuals can suspend use of one or more features of the application. A tool that can accommodate intermittent use raises questions of how to reengage the user at timely moments when repeatedly abnormal results or new trends are detected. If designed properly and flexibly, such technologies could provide patients with the support that is most important to them at their own stage of illness, communicate with their providers more effectively, and maximize the technology’s effectiveness in helping patients improve their self-management and their health.

Acknowledgments

None

References


Evolving Patient Compliance Trends: Integrating Clinical, Insurance, and Extrapolated Socioeconomic Data

Joseph J. Klobusicky, PhD, Arun Aryasomayajula, Nicholas Marko, MD
Geisinger Health System, Danville, PA

Abstract

Efforts toward improving patient compliance in medication focus on either identifying trends in patient features or studying changes through an intervention. Our study seeks to provide an important link between these two approaches through defining trends of evolving compliance. In addition to using clinical covariates provided through insurance claims and health records, we also extracted census based data to provide socioeconomic covariates such as income and population density. Through creating quadrants based on periods of medicine intake, we derive several novel definitions of compliance. These definitions revealed additional compliance trends through considering refill histories later in a patient’s length of therapy. These results suggested that the link between patient features and compliance includes a temporal component, and should be considered in policymaking when identifying compliant subgroups.

Introduction

In the field of pharmaco economics, compliance (or adherence) to a medication is defined as the degree to which a patient follows instructions from a health care provider\(^1\). Compliance measures span over all fields of treatment, including preventative measures such as exercise and diet. However, the most intensely studied form of compliance is with prescription drug consumption, due to its major economic impact. Approximations for annual costs of noncompliance due to medication related hospital visits are as high as $100 billion\(^2,3,4,5\), while the consulting group Capgemeni Worldwide recently published the striking estimate of $564 billion of annual costs suffered globally by pharmaceutical companies, of which American companies lose $188 billion\(^6\). Along with staggering economic losses, patient compliance also presents a major hurdle to patient health. For instance, compliance rates below 90\% for HIV patients can cause viral replication and disease progression\(^7,8\), while for diabetics, proper compliance is essential in preventing hypertension and myocardial infarction\(^9\).

Given a specific prescription, compliance can be further classified with respect to the potential ways a patient can deviate from a provider’s instructions. Primary compliance is defined as a patient’s fidelity of filling and refilling prescriptions. Secondary compliance refers to whether a patient actually consumes their medication. While direct observation and blood tests can validate secondary compliance, these methods are mostly infeasible, and instead rely on methods such as surveys\(^1\). Our data focuses on an aspect of primary compliance which observes patterns of drug intake conditioned on the existence of a patient refill history. Our concern, therefore, is not the event that a patient has taken a medication, but rather the discrepancy between recommended and observed fill dates.

Studies based on claims data most commonly use \(\text{MPR}\), or medication possession ratio, as a means for quantifying patient compliance. This ratio is defined as the total number of medication supply days divided by the total length of therapy. To illustrate, a patient with a prescription history spanning 60 days receiving 2 prescription fills for 15 days is assigned an \(\text{MPR}\) of 30/60, or .50. Studies have linked high \(\text{MPR}\) with commonly used covariates such as age, race, and sex\(^10\), as well as insurance related quantities including out-of-pocket payments\(^11,12\). Our study diverges significantly from other investigations through a novel analysis of evolving patient trends. Rather than testing changes in compliance under some intervention (e.g. smart pill counters or medication reminders), we define notions of general, initial, eventual, and consistent compliance, and compare patient trends across these different features. These notions of compliance can address questions of policy, which seek not only to identify high risk patients, but also those patients most likely to respond to intervention.

Our main dataset is insurance claims data of prescription orders. While not accounting for secondary compliance, fill records provide an objective history of intake that is resistant to biased responses, which can arise in survey data. Insurance claim data is natural for questions concerning consistency of medication refills, as the dataset includes prescription fill dates, supply lengths, and refill numbers. This approach was taken by Soloman and colleagues\(^13\) and Gibson and colleagues\(^14\) for studying the relation of cost sharing and compliance. Our statistical approach is similar to Rolnick and colleagues\(^10\), who also use prescription fill data to examine trends, including census derived
covariates of education, income, and poverty. For our study, we use extrapolated data from the US Census and American Community Survey (ACS) to examine compliance with median family income and population density.

Figure 1: Data flow. A data broker combines GHP (insurance) and EHR (clinical) datasets, available from GHS database (lower left). Zip codes from the EHR dataset are used to obtain extrapolated census data (top left). The data broker (center) then deidentifies all data for statistical analyses (right).

Data Sources

All clinical and insurance claims data is aggregated through available databases of Geisinger Health System (GHS), a hospital located in central Pennsylvania which serves approximately 2.6 million people, mostly residing in surrounding rural areas. To collect clinical features such as body mass index (BMI), we restrict attention to patients who have both insurance and clinical data. The use of city-based geographic data, an instance of PHI (Protected Health Information), in addition to other deidentified data for this study was approved through an IRB.

Using zip code information from EHR data, we use census data to create an enhanced dataset containing socioeconomic factors, with coarseness at the city level. For the purposes of anonymization, we use the following scheme. A certified data broker deidentifies GHP and EHR data through random date shifting and random patient number identifier assignments. All PHI sensitive data is restricted to a work computer inaccessible to other investigators. From this computer, the data broker collects web-based data, and returns deidentified data, along with socioeconomic data, to an analyst for running statistical models. An illustration of the data flow between agents is given in Figure 1.

The chief dataset which contains information about prescription fills is collected through the Geisinger Health Plan (GHP), a subsidiary HMO of GHS. For each patient in the GHP database, there exists a collection of prescriptions taken, each with its separate history. For each of these drugs there are lists of dates corresponding to refill occurrences and supply length of medication. As the purpose of this study is to investigate changes in compliance over a series of refills, we will require at least three fills and a minimum of 180 day length of therapy. Each separate prescription for a person taking multiple prescriptions will be treated individually. Thus, in the sequel, to avoid confusion, we will use the term compliance of prescriptions, rather than compliance of patients, to denote scores for a single prescription. Through taking averages over all prescriptions, we can calculate patient compliance scores in a straightforward manner.

For a prescription with $N$ refill dates $d_1, \ldots, d_N$, with corresponding supply of prescription $s_1, \ldots, s_N$ given in days, we can define the medical possession ratio ($MPR$) as

$$MPR = \frac{\sum_{i=1}^{N-1} s_i}{|d_N - d_1|}.$$
where $|d_N - d_1|$ denotes the length in days between $d_N$ and $d_1$. Note that the final fill is not taken into account when calculating $MPR$, as we have no way of deducing a patient’s compliance if there is no subsequent refill. We will use the conventional definition of a compliant prescription as one with an $MPR$ of greater than .80, a commonly used threshold.

Electronic health records (EHR) were obtained from patients who were also members of GHP. The entirety of data available from EHR is massive in both magnitude and variety, containing millions of records of patient medications, events such as surgeries, complications such as occurrence of disease, and lengths of stay. While this rich dataset is certainly of future interest for more specific investigations, for this study we have selected a restricted set of covariates, consisting of age, sex, race, and BMI. Socioeconomic data was provided through the 2010 US Census and 2006-2010 ACS, which provide specific socioeconomic features. For our additional variables, we have chosen median income and population density.

Methods:

Definitions of Evolving Compliance. We now define different ways of describing compliance based on changes in behavior in time. To do so, we separate each prescription with a length of therapy of $L$ days into an initial period of day range $r$ to $\lfloor L/2 \rfloor$, and a final period of day range of $\lceil L/2 \rceil + 1$ through $L$, where $[x]$ denotes the floor value of $x$. For each period and prescription, we calculate the corresponding $MPR$ to give us initial and eventual compliance scores. This split defines each prescription as belonging to one of four quadrants Q1-Q4:

**Q1: Consistently Compliant.** A prescription is compliant in both initial and final periods.
**Q2: Compliant to Noncompliant.** A prescription is compliant in the initial period, and noncompliant in the final period.
**Q3: Consistently Noncompliant.** A prescription is noncompliant in both initial and final periods.
**Q4: Noncompliant to Compliant.** A prescription is noncompliant in the initial period, and compliant in the final period.

There are seven ways of dividing quadrants, either through comparing one quadrant against three quadrants, or two quadrants against another two quadrants. Our study focuses on three such divisions with conditions (see Figure 2):

1. A prescription is **consistently compliant** (Q1 vs. Q2, Q3, and Q4).
2. A prescription is **initially compliant** (Q1 and Q2 vs. Q3 and Q4).
3. A prescription is **eventually compliant** (Q1 and Q4 vs. Q2 and Q3).

We refer to a patient who is compliant over the entire length of therapy as **generally compliant**. Thus, each prescription was given four classifications based on which types of compliance are realized.

Statistical Design. For each variety of compliance, we ran chi-square tests of independence and logistic regression models through binarizing inputs. For EHR, we classified prescriptions as white, male, obese ($BMI >30$), or above median age (54 years old). For census data we classified by median household income ($\$43,837$ per year) and population density (245 residents per square mile). We repeat the fact that individual prescriptions were taken as separate data points, meaning that a patient with $N$ prescriptions provided $N$ separate data points with repeated inputs and varied outputs of $MPR$. The full dataset consists of 75940 prescriptions for 8889 patients. The median $MPR$ is 65.5%. This number is smaller than other studies which take $MPR$ as a compliance measure for specific chronic conditions. However, our study was over the space of all prescriptions, including those which were non-chronic or with more irregular drug intake patterns.
Figure 2: Variations of compliance. (a) A prescription record over a length of therapy consists of an initial first period and subsequent second period. Tiles denote individual days, and prescription symbols define fills. (b) Considering $MPR$ only over the entire length of therapy, patients can be partitioned as compliant or noncompliant based on a threshold score of .80. (c) Left: Considering $MPR$ over both first and second periods, patients can fall into one of four types of compliance behaviors. Right: Separating quadrants creates several definitions of compliance, including compliance over first period (initial), second period (eventual), and both periods (consistent).

Results

Summaries of chi-square tests for independence and multivariate logistic regression models are presented for the four compliance types in Tables 1-4 of the Appendix. The $p$ values for logistic regression are based on a Wald test with the null hypothesis that odds ratios are equal to 1. Age and race are the strongest predictors of compliance, as white and older patients have higher scores across all definitions of compliance. While race has a greater difference of compliance scores than age, $p$ values are consistently lower, as the sample population was largely homogenous (about 90% white). Males and obese patients are also more likely to have higher compliance scores, although differences in compliance are less pronounced than those for age and race. Census related data, income and density, produce the smallest differences in compliance, suggesting that high income and low population density related to higher compliance. However, while compliance scores are slightly higher for those living in more rural areas, we are not able to reject the null hypothesis that population density is independent of compliance.

Consistent and eventual compliance produce the most statistically significant variables for $p<.05$, whereas initial compliance produce the least. In particular, a small, but statistically significant, difference in scores was reported with respect to income data when observing consistent and eventual compliance, while there no such evidence of a difference with initial and general compliance. Also, race produces statistically significant differences in all definitions of compliance except for initial compliance.
Discussion

Changing Compliance Measures Imply Changing Significant Populations

While we observed certain covariates as statistically significant in terms of general compliance, the focus of this study was to develop alternative metrics which can capture information not available from traditional measures of compliance. In this analysis, we create variations in scores by considering compliance as a function dependent upon time. When considering compliance over the entire length of therapy for a patient, the more random nature of initial compliance can hide trends that more clearly occur after a patient has received several refills. A direct method of removing initial noise would consider either consistent or eventual compliance. While consistent compliance does, in fact, capture new socioeconomic trends, requirements are more stringent, leading to a smaller set of testable patients. Thus, it is reasonable to condition medication compliance scores after several refills have taken place, especially in the case of studying chronic conditions spanning over several years. The question of how many refills we should exclude before testing would be specific to the type of condition of question, both in the inherent compliance statistics, as well as the relative important for early compliance to medication.

New perspectives of analyzing quadrants can also shed light on different topics of interest. One particularly interesting approach is to investigate behavior in quadrants Q2 and Q4, which describe patients who experience a change in compliance. In terms of policy, a practical option when considering target audiences might focus on those who are susceptible to changes. While our current set of covariates was too generic to provide a definitive profile of such patients, we have provided a metric that can measure such subgroups for new studies with more specific datasets.

New Directions for Compliance Metrics

Our extensions to medication compliance can be used for providing a notion of “persistence” for claims related studies, which is considered as a distinct measure. Medication persistence is defined as the act of continuing a treatment for a prescribed duration. While claims data only offers information on prescriptions which have been filled, notions such as consistent compliance may be seen as a type of persistence with respect to the lag between drug consumption and refill dates. Notions of initial and eventual compliance should also be taken into account when conducting studies which focus on temporal changes, as the population may have tendencies toward compliance change. We also note that this type of analysis is an initial step for targeted intervention of patients who are likely to experience changes in compliance.

A natural extension of our study would consider MPR as a continuous function of time. One such approach for comparing MPR can be understood through comparing cumulative functions of total missed medication days. Through an appropriate normalization, we can then compare cumulative distributions for different prescription histories. While the common method for comparing such functions would be through calculating a Kolmogorov-Smirnov distance, the promising field of quantile regression, popular in econometrics, offers more relevant information on quantiles conditioned on certain feature sets. While studies have used methods from quantile regression, they use general compliance rates as the object of study, without considering variations in time. Another point of research would examine the differences between adherence and a patient’s medication portfolio, including focusing on frequency of intake effects on coupling certain drugs.

Another future direction of study involves the almost limitless source of data available from text scraping or data aggregation services. Strategies of utilizing social media tend toward intervention, where findings are currently mixed. A comprehensive study from Scheurer and colleagues of over 5000 articles reached the conclusion that online social support networks, including social media, can increase compliance. However, the same study showed that “practical social support”, such as offering transportation to pharmacies, offered the greatest increase. A less studied aspect uses alternative data sources for compliance prediction. Examples of success stories are from social feeds such as Twitter and Yelp restaurant reviews, which have been used to track the progression of influenza and post-partum depression. For socioeconomic issues, data aggregators can provide more specific public information which may relate to compliance, such as housing and criminal justice records. This additional inclusion of data may allow for a more advanced analysis of compliance scores through machine learning techniques.
Conclusion

Estimates of patient compliance are sensitive to temporal considerations. Trends toward eventual compliance are similar to those of overall compliance, while compliance near the beginning of a prescription tends to be more uniformly random over all features. This suggests that studies should consider conditioning compliance studies after a patient receives several refills. Future work will focus on more extrinsic data collection and partitioning methods that more accurately capture compliance trends.

References


Appendix: Tables of Results

Table 1: Statistics for General Compliance. Chi-square and multivariate logistic regression tests are performed with respect to general compliance, or an MPR ≥ .80 over length of therapy. Odd ratios are computed for the first vs. second entry in each category (age ≥ 54 years vs. age < 54 years, white vs. nonwhite, etc.) against probabilities of compliance.

<table>
<thead>
<tr>
<th>Chi-Square Test for Independence</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>% Compliant</td>
</tr>
<tr>
<td>Age ≥ 54 years</td>
<td>39.23***</td>
</tr>
<tr>
<td>Age &lt; 54 years</td>
<td>32.50</td>
</tr>
<tr>
<td>White</td>
<td>35.60**</td>
</tr>
<tr>
<td>Nonwhite</td>
<td>26.35</td>
</tr>
<tr>
<td>Obese (BMI ≥ 30)</td>
<td>37.20*</td>
</tr>
<tr>
<td>Not Obese</td>
<td>34.98</td>
</tr>
<tr>
<td>Male</td>
<td>37.04**</td>
</tr>
<tr>
<td>Female</td>
<td>34.24</td>
</tr>
<tr>
<td>Family Income ≥ $43,837 per year</td>
<td>35.59</td>
</tr>
<tr>
<td>Family Income &lt; $43,837 per year</td>
<td>34.57</td>
</tr>
<tr>
<td>Pop. Density ≥ 245 per sq. mile</td>
<td>34.61</td>
</tr>
<tr>
<td>Pop. Density &lt; 245 per sq. mile</td>
<td>35.51</td>
</tr>
</tbody>
</table>
Table 2: **Statistics for Consistent Compliance.** Chi-square and multivariate logistic regression tests are performed with respect to consistent compliance, or an $MPR \geq .80$ over both first and second periods of therapy. Odd ratios are computed for the first vs. second entry in each category (age $\geq 54$ years vs. age $< 54$ years, white vs. nonwhite, etc.) against probabilities of compliance.

<table>
<thead>
<tr>
<th>Feature</th>
<th>% Compliant</th>
<th>Logistic Regression</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age $\geq 54$ years</td>
<td>25.84***</td>
<td>Age</td>
<td>.73***</td>
</tr>
<tr>
<td>Age $&lt; 54$ years</td>
<td>19.39</td>
<td>Race</td>
<td>.56***</td>
</tr>
<tr>
<td>White</td>
<td>22.36***</td>
<td>BMI</td>
<td>.80***</td>
</tr>
<tr>
<td>Nonwhite</td>
<td>12.84</td>
<td>Sex</td>
<td>.86**</td>
</tr>
<tr>
<td>Obese (BMI $\geq 30$)</td>
<td>23.88*</td>
<td>Income</td>
<td>.88*</td>
</tr>
<tr>
<td>Not Obese</td>
<td>21.68</td>
<td>Density</td>
<td>.95</td>
</tr>
<tr>
<td>Male</td>
<td>23.52***</td>
<td>***: p value &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>21.14</td>
<td>** : p value &lt; .01</td>
<td></td>
</tr>
<tr>
<td>Family Income $\geq 43,837$ per year</td>
<td>22.69*</td>
<td>* : p value &lt; .05</td>
<td></td>
</tr>
<tr>
<td>Family Income $&lt; 43,837$ per year</td>
<td>20.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop. Density $\geq 245$ per sq. mile</td>
<td>21.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop. Density $&lt; 245$ per sq. mile</td>
<td>22.18</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: **Statistics for Initial Compliance.** Chi-square and multivariate logistic regression tests are performed with respect to initial compliance, or an $MPR \geq .80$ over the first period of therapy. Odd ratios are computed for the first vs. second entry in each category (age $\geq 54$ years vs. age $< 54$ years, white vs. nonwhite, etc.) against probabilities of compliance.

<table>
<thead>
<tr>
<th>Feature</th>
<th>% Compliant</th>
<th>Logistic Regression</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age $\geq 54$ years</td>
<td>52.71***</td>
<td>Age</td>
<td>.83***</td>
</tr>
<tr>
<td>Age $&lt; 54$ years</td>
<td>47.12</td>
<td>Race</td>
<td>.86</td>
</tr>
<tr>
<td>White</td>
<td>49.67</td>
<td>BMI</td>
<td>.85***</td>
</tr>
<tr>
<td>Nonwhite</td>
<td>44.59</td>
<td>Sex</td>
<td>.88**</td>
</tr>
<tr>
<td>Obese (BMI $\geq 30$)</td>
<td>51.37*</td>
<td>Income</td>
<td>1.01</td>
</tr>
<tr>
<td>Not Obese</td>
<td>49.08</td>
<td>Density</td>
<td>.99</td>
</tr>
<tr>
<td>Male</td>
<td>51.43**</td>
<td>***: p value &lt; .001</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>48.34</td>
<td>** : p value &lt; .01</td>
<td></td>
</tr>
<tr>
<td>Family Income $\geq 43,837$ per year</td>
<td>49.54</td>
<td>* : p value &lt; .05</td>
<td></td>
</tr>
<tr>
<td>Family Income $&lt; 43,837$ per year</td>
<td>49.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop. Density $\geq 245$ per sq. mile</td>
<td>49.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop. Density $&lt; 245$ per sq. mile</td>
<td>48.97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4: **Statistics for Eventual Compliance.** Chi-square and multivariate logistic regression tests are performed with respect to consistent compliance, or an $MPR \geq .80$ over the second period of therapy. Odd ratios are computed for the first vs. second entry in each category (age $\geq 54$ years vs. age $< 54$ years, white vs. nonwhite, etc.) against probabilities of compliance.

<table>
<thead>
<tr>
<th>Chi-Square Test for Independence</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feature</strong></td>
<td><strong>% Compliant</strong></td>
</tr>
<tr>
<td>Age $\geq 54$ years</td>
<td>47.82***</td>
</tr>
<tr>
<td>Age $&lt; 54$ years</td>
<td>41.98</td>
</tr>
<tr>
<td>White</td>
<td>44.54**</td>
</tr>
<tr>
<td>Nonwhite</td>
<td>36.15</td>
</tr>
<tr>
<td>Obese (BMI $\geq 30$)</td>
<td>46.29*</td>
</tr>
<tr>
<td>Not Obese</td>
<td>43.86</td>
</tr>
<tr>
<td>Male</td>
<td>44.99</td>
</tr>
<tr>
<td>Female</td>
<td>43.81</td>
</tr>
<tr>
<td>Family Income $\geq$ $43,837$ per year</td>
<td>45.04*</td>
</tr>
<tr>
<td>Family Income $&lt; $43,837$ per year</td>
<td>42.62</td>
</tr>
<tr>
<td>Pop. Density $\geq 245$ per sq. mile</td>
<td>44.88</td>
</tr>
<tr>
<td>Pop. Density $&lt; 245$ per sq. mile</td>
<td>44.10</td>
</tr>
</tbody>
</table>

***: p value < .001
** : p value < .01
*  : p value < .05
Automated Reconciliation of Radiology Reports and Discharge Summaries

Bevan Koopman, PhD1,2, Guido Zuccon, PhD2, Amol Wagholikar, PhD1, Kevin Chu, MBBS FACEM3, John O'Dwyer1, Anthony Nguyen, PhD2, Gerben Keijzers, MBBS PhD FACEM4

1Australian e-Health Research Centre, CSIRO, Brisbane, QLD, Australia; 2Queensland University of Technology, Brisbane, QLD, Australia; 3Royal Brisbane and Women’s Hospital, Brisbane, QLD, Australia; 4Gold Coast Hospital, Gold Coast, QLD, Australia.

ABSTRACT

We study machine learning techniques to automatically identify limb abnormalities (including fractures, dislocations and foreign bodies) from radiology reports. For patients presenting to the Emergency Room (ER) with suspected limb abnormalities (e.g., fractures) there is often a multi-day delay before the radiology report is available to ER staff, by which time the patient may have been discharged home with the possibility of undiagnosed fractures. ER staff, currently, have to manually review and reconcile radiology reports with the ER discharge diagnosis; this is a laborious and error-prone manual process. Using radiology reports from three different hospitals, we show that extracting detailed features from the reports to train Support Vector Machines can effectively automate the identification of limb fractures, dislocations and foreign bodies. These can be automatically reconciled with a patient’s discharge diagnosis from the ER to identify a number of cases where limb abnormalities went undiagnosed.

Introduction

The misdiagnosis of a patient’s true clinical condition due to misinterpretation of radiological evidence by the treating clinician is an occasional problem in hospital emergency departments. There is always a time delay between reporting of the radiologist and clinical treatment by the Emergency Room (ER) clinician. The large amount of manual processing of unstructured text is one of the main issues that can be resolved by technology enabled solutions.

A good example of a misdiagnosis issue is the identification of subtle limb abnormalities (fractures, dislocation or foreign bodies). Radiological evidence of limb abnormalities, when subtle, can be missed by clinicians working in the ER. The reporting of a abnormalities by a radiologist may not occur in real time and therefore may not be available to the clinician treating a patient. Consequently, patients may be sent home without appropriate treatment and follow up. A study by Cameron1 reported that 2.1% of all fractures were not identified on their initial presentation to the ER. Furthermore, Sprivulis and Frazer9 reported that 1.5% of all x-rays have abnormalities not identified in the ER records. Similarly, Mounts et al.5 reported that 5% and 2% of the x-rays of the hand/fingers and ankle/foot from a paediatric ER had fractures missed by the treating clinician. Although small, these percentages are not insignificant.

The need to reduce errors is well recognised4,7,8. To ensure a diagnosis is not missed, radiology reports are commonly checked and patient records are reviewed, but this may not happen until days after the initial presentation. The current clinical practice of identifying limb abnormalities from radiology reports is highly labour intensive and is subject to human error or omissions. There is a need to streamline the process of identifying missed abnormalities for better patient outcomes. Technology enabled solutions that can streamline the diagnosis identification would certainly improve efficiency in the existing process.

Previous work has focused on automatically detecting fractures from free-text radiology reports. De Bruijn et al.3 considered acute fractures of the wrist and reported that a Support Vector Machine algorithm (SVM) was able to identify fractures in free-text radiology notes, achieving an overall F-measure of 91.3%. While, Thomas et al.10 developed a text search algorithm that accurately classified radiology reports into the categories "fracture", "normal" and “neither normal nor fracture”. Zuccon et al.15 have studied Naive Bayes and Support Vector Machines based classifiers for the identification of limb abnormalities. They have shown that machine learning techniques coupled with both word and semantic features are very effective for this task, achieving an overall F-measure of 92.3%. Their evaluation however was limited to a sample of 99 radiology reports from a single hospital radiology service.
Table 1: Three different datasets of radiology reports, along with the number of normal and abnormal cases as identified through our annotation process. The average document length for free-text reports in each dataset is also recorded: the large difference in average length between GCH and RBWH/RCH may be due to differences in reporting language and style conventions.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>#Reports</th>
<th>Normal</th>
<th>Abnormal</th>
<th>Avg. Doc. Len.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBWH</td>
<td>Royal Brisbane &amp; Womens’ Hospital (adult)</td>
<td>1,480</td>
<td>58%</td>
<td>42%</td>
<td>52 words</td>
</tr>
<tr>
<td>RCH</td>
<td>Royal (Brisbane) Childrens’ Hospital (child)</td>
<td>498</td>
<td>66%</td>
<td>34%</td>
<td>50 words</td>
</tr>
<tr>
<td>GCH</td>
<td>Gold Coast Hospital (adult child 38%)</td>
<td>400</td>
<td>62%</td>
<td>38%</td>
<td>27 words</td>
</tr>
</tbody>
</table>

In this paper, we build upon the work of Zuccon et al.\textsuperscript{15} and we experiment with the automatic classification of free-text radiology reports for identifying abnormalities of limb structures using machine learning algorithms and features such as bigrams formed by stemmed tokens, negations, and SNOMED-CT concepts extracted from the free-text. While previous work has shown promise, it does not address a number of important areas for the practical use of such techniques in a clinical setting. We outline these below and highlight the contribution this study makes in addressing each.

1. Where the majority of the existing methods only used term-based features we extract medical concepts from the SNOMED-CT ontology and exploit these for training the classification model.

2. Previous studies used datasets that were small and homogenous\textsuperscript{10,3,15}. Instead, in this study, we use a larger set of reports taken from different hospitals and different ages (adults vs. children). An important requirement for the general applicability of these methods is the robustness of the models across different hospital datasets; i.e., how well would one method, developed using one hospital’s set of reports work when deployed at another hospital, where particular institutional conventions may result in different language and authoring styles in radiology reports. In addition, radiology reports for children and adults may also differ. To address this issue, we evaluate our method on three different sets of radiology reports from three different hospitals (adults and children). These types of heterogeneous data sources were not considered in previous studies. In this study, we empirically show, via different training/test combinations, that the method developed on one hospital’s reports (with differing conventions and patient cohorts) can be applied to another with marginal loss in effectiveness.

3. Finally, and perhaps most importantly, we investigate how the classification of radiology reports can be used in a real clinical setting to reconcile the radiology diagnosis with that of the patient’s discharge diagnosis from the ER, thereby identifying a number of cases where limb abnormalities may have been undiagnosed. Thus we study an end-to-end application of natural language processing and machine learning to aid clinicians in the identification of undiagnosed limb abnormalities.

Materials and Methods

Data

A set of 2,378 free-text radiology reports of limb structures was acquired from the Emergency Department of three large Australian public hospitals (adult, children and mixed adult/children). Ethics approval was granted by the Human Research Ethics Committee at Queensland Health to use the non-identifying data. Free-text reports were short in length, containing on average 47 words, and an (unstemmed) vocabulary comprising 4846 unique words. Details of the three datasets are outlined in Table 1. Free-text reports in the GCH dataset were found to be on average consistently shorter than those in the other two datasets: this may indicate differences in reporting style and conventions between the hospital sites.

Free-text reports were manually annotated by an Emergency Medicine Registrar and a Medical Officer as being either: “normal” — the radiography does not exhibit a fracture, dislocation or presence of a foreign body; or
“abnormal” — some fracture, dislocation or foreign body was found.

A software tool was developed to assist clinicians in the recording of their interpretation and to highlight the portion of text in the report that lead to their interpretation.

Initially, assessors agreed on the annotations of 2,215 out of the 2,378 reports. A senior Staff Specialist in Emergency Medicine was then asked to act as third assessor and resolve disagreements. The distribution of normal and abnormal cases across the three datasets is reported in Table 1. The Fleiss’ kappa ($\kappa$) calculated on the initial set of annotations provided by the two first assessors was 0.85, thus exhibiting strong inter-rater reliability.

### Automatic Feature Extraction and Weighting

Machine learning algorithms require documents to be described by features. The text analysis capabilities of the Medtex tool were developed to automatically extract features from the free-text radiology reports. Medtex is a text analysis system that has been previously used for classify cancer-notifiable pathology reports and produce a minimum set of synoptic factors. A wide range of features were initially extracted, including:

- token, i.e., a word found in a report;
- punctuation;
- token stem, i.e., the stemmed version of a word contained in a report;
- token negation, i.e., if a token or phrase was explicitly negated (e.g., “no fracture”); the Medtex implementation of the ConText algorithm was used to identify negations in free-text;
- token stem bi-gram, i.e., a pair of adjacent stemmed words as found in a report;
- token stem tri-gram, i.e., a 3-tuple of adjacent stemmed words contained in a report;
- SNOMED-CT concepts extracted from the text of the report;
  - the fully specified terms of extracted SNOMED-CT concepts restricted to morphologic abnormalities and disorders;
  - SNOMED-CT concept bi-gram, i.e., a pair of adjacent SNOMED-CT concepts as found in a report.

While a number of these features are commonly used for the classification of free-text documents, the use of SNOMED-CT features have not been widely evaluated by previous works on classification of radiology reports. To our knowledge, only Zuccon et al. investigated these features but their evaluation was limited to a small sample of 99 radiology reports. In this work, SNOMED-CT concepts were extracted by annotating the radiology reports with the MetaMap: a natural language processing tool that identifies medical concepts mentions in free-text. The actual feature used in training the classifier was the SNOMED-CT concept ids found my MetaMap. Previous empirical results have shown that SNOMED-CT concepts, in particular those referring to abnormalities (e.g., fracture, dislocation, etc.) and disorders (e.g., fracture of bone, traumatic injury, etc.), provide valuable evidence for representing free-text radiology report data. Table 2 provides an example of feature sets extracted from the free-text of the radiology reports.

<table>
<thead>
<tr>
<th>Features</th>
<th>stem</th>
<th>Features</th>
<th>stemBigram</th>
<th>concept</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Report 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Automatic Classification and Evaluation Methodology

To classify radiology reports we used the Weka toolkit API and the corresponding implementation of the Sequential Minimization Optimization (SMO) classifier. The SMO classifier is a support vector machine (SVM) algorithm where training is performed according to the sequential minimal optimisation algorithm and a polynomial kernel is used. The parameters of all classifiers were set to the default values (see Witten et al. for details).

To explore the effectiveness of the machine learning classifier for identifying limb abnormalities we conducted three sets of experiments.

Experiment 1. We combine all three datasets of Table 1 and use the 10-fold cross validation methodology to evaluate the classification algorithms. In this methodology, the dataset is randomly divided into 10 stratified folds of equal dimension (in our case nine folds will contain 238 reports, while the remaining fold will contain only 236 reports). The model for each classifier was then learnt on nine of these folds, leaving one fold out for testing the model. The process was repeated by selecting a new fold for testing, while a new model was learnt from the remaining folds. Classification performances were then averaged across the folds left out in each iteration. The aim of this experiment was to evaluate the effectiveness of a classifier learnt on the whole combination of datasets.

Experiment 2. We consider the reports from each hospital dataset separately. For each dataset, we use the 10-fold cross validation methodology to evaluate the classification algorithms specific to that dataset. Thus, experiments on the larger RBWH dataset are characterised by larger folds than those on the remaining two smaller datasets. The aim of this experiment was to evaluate the effectiveness of classifiers specifically learnt on individual datasets and thus individuate whether a dataset is more challenging than others for automatic classification (and what the possible causes for this are).

Experiment 3. We performed a split train/test evaluation: train on one hospital’s reports and test on another hospital’s reports. This procedure was repeated for all combinations of hospitals and included training on reports from two hospitals and testing on those from another. The aim of this last experiment was to evaluate the robustness of our method across reports from different hospitals.

As a baseline for comparison against our machine learning method, we included a keyword spotting system, which resembled the method by Thomas et al. and Wagholkar et al. A set of regular expressions were defined based on common phrases or terms that identify an abnormality; these were based on discussion with a senior Staff Specialist in Emergency Medicine. (Details of the regular expressions are provided in Appendix.)

Two evaluation measures were considered: precision and recall (also called positive predictive value and sensitivity, respectively). Precision is the fraction of positively classified reports that contain abnormalities, while recall is the fraction of actual abnormalities that were positively classified. In addition, to provide a single, overall evaluation measure, precision and recall are combined into a third evaluation measure, F-measure.

Reconciliation of Radiology Reports with Emergency Room Discharge Diagnosis

Using the methods described here we are able to automatically identify abnormalities from a patient’s radiology report. The benefit of such a method is the ability to reconcile the abnormality classification with the discharge diagnosis from the ER to ensure that an abnormality did not go unrecognised and the patient discharged without proper treatment. To demonstrate the utility of this, we reconciled all the radiology reports used in the classification task with the ER discharge diagnosis ICD-10 code. If the ER discharge diagnosis ICD-10 code matched a predefined set of “abnormal” codes then the patient was marked as abnormal; else they were marked “normal”.* (The full list of ICD-10 codes considered as abnormal was provided by an ER clinician (KC) and is provided in Appendix.) Patients that had a abnormal radiology classification using our automated method but did not have any abnormality related ICD-10 code recorded in the ER discharge diagnosis were flagged as possible misdiagnosis for immediate followup.

*Note that in Australian Emergency Departments ICD-10 codes are used as a diagnostic classification and are not used for billing purposes.
Table 3: Classification results for each of the three datasets, comparing the proposed machine learning method (SVM) against the keyword baseline. The percentage change in F-measure shows the improvement of SVM over the keyword baseline.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBWH</td>
<td>Keyword</td>
<td>0.73</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.88</td>
<td>0.91</td>
<td>0.89 (+33%)</td>
</tr>
<tr>
<td>RCH</td>
<td>Keyword</td>
<td>0.74</td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95 (+25%)</td>
</tr>
<tr>
<td>GCH</td>
<td>Keyword</td>
<td>0.94</td>
<td>0.58</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95 (+32%)</td>
</tr>
<tr>
<td>All</td>
<td>Keyword</td>
<td>0.76</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92 (+33%)</td>
</tr>
</tbody>
</table>

Table 4: Split dataset training/testing F-measure results. Grey shaded cells represent cross validation results; all other results are train/test.

<table>
<thead>
<tr>
<th>Testing</th>
<th>GCH</th>
<th>RBWH</th>
<th>RCH</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCH</td>
<td>0.95</td>
<td>0.80</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>RBWH</td>
<td>0.84</td>
<td>0.89</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>RCH</td>
<td>0.87</td>
<td>0.81</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>GCH+RBWH</td>
<td></td>
<td></td>
<td></td>
<td>0.88</td>
</tr>
<tr>
<td>GCH+RCH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBWH+RCH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>0.84</td>
<td>0.80</td>
<td>0.92</td>
<td></td>
</tr>
</tbody>
</table>

Classification Results and Discussion

The overall classification results (Experiments 1 and 2) are shown in Table 3. F-measure was used as the overall effectiveness measure and the percentage change in F-measure shows the improvement of the machine learning method over the keyword baseline. For all hospital datasets, the SVM method outperformed the keyword baseline in all evaluation settings. This is consistent with previous results. In addition, the keyword baseline had more variance across datasets (lower on RBWH, higher on RCH), whereas the SVM method was more stable across datasets. For the SVM method, both precision and recall were of similar value, indicating that the errors that did occur were a mixture of false positive and false negatives.

Split Dataset Training and Testing

The split training/testing F-measure results (Experiment 3) are shown in Table 4. The first column in the table corresponds to the dataset used in training; the first row corresponds to the dataset the model was tested on; grey shaded cells are the previous cross validation results repeated from Table 3. Comparison with the grey cell cross validation results shows that there was some loss in effectiveness when applying models across datasets (e.g., F-measure of RBWH was 0.89 for cross validation and 0.80–0.81 when trained on other datasets). However, even with some loss in effectiveness, the results indicate that the models were still robust when applied across datasets (certainly compared to the results of the keyword baseline).

Combining datasets for training (e.g., train on GCH+RBWH, test on RCH) showed no significant benefit over using a single dataset to train (e.g., train on RBWH, test on RCH). This result reveals that simply adding more training data
Figure 1: Comparison of similarity between datasets (or itself) and F-measure effectiveness of the classifier. Points with two datasets represents a different train/test combinations RCH/GCH = train on RCH test on GCH), while points with a single dataset (e.g., RCH) are the cross validation results for a single dataset.

does not lead to immediate improvement in effectiveness. Instead, effectiveness was more influenced by the particular dataset used for training. For example, when testing on GCH, it was better to train on RCH than RBWH, even though RBWH was a larger dataset. Based on this finding we set out to understand how similar each of the three datasets were to each other in order to explain the differences in effectiveness in training/testing combinations.

Dataset Similarity and its Effect on Performance

A pairwise similarity calculation was made between the three datasets. This was done by comparing the similarity of every document in one dataset to every other document in another dataset and recording the overall mean similarity. The similarity measure between two individual documents can be calculated by taking the cosine angle between the two documents’ term vectors\[†\] — a standard approach applied in information retrieval when comparing text\[11\]. Note that the average similarity of a dataset to itself (e.g., GCH vs. GCH) can actually be interpreted as a cohesiveness measure: how similar reports in the dataset are to each other. To understand the similarity results in light of the classifier effectiveness we provide a plot of similarity vs. F-measure in Figure 1 and discuss this in further detail below.

We first consider the cohesiveness of each dataset with itself (i.e., single dataset points). The child-only reports of RCH were the most cohesive, while the mixed adult/child reports of GCH were understandably the least cohesive; adult reports from RBWH were in between. The cross validation training on these datasets obtained the best F-measure — obviously it is best to train and test on same dataset. For these three cross-validated datasets, similarity did not correlate with F-measure (e.g., GCH still had a high F-measure but the lowest similarity score).

Comparing across datasets, any combination containing both RCH (children) and GCH (mixed) had the lowest similarity yet produced the best F-measure (and effectiveness was similar for both RCH/GCH and GCH/RCH). In contrast, for the other four dataset combinations that contained RBWH, the F-measure was lower. In addition, there was a large difference in F-measure for swapping the training/test combination, i.e., RBWH/RCH was lower than RCH/RBWH. For all four combinations that involved RBWH, training on RBWH was better than testing on RBWH. An immediate

\[†\]A term vector \(\vec{d}\) for document \(d\) is an \(n\) dimensional vector, where \(n\) is the size of the (stemmed) vocabulary. Each element in \(\vec{d}\) is the TF-IDF weight of a stemmed term from the vocabulary in the document \(d\).
Reconciliation Results and Discussion

The reconciliation process involved checking the classification of a patient’s radiology report with their ICD-10 discharge diagnosis from the ER. Four different ER / Radiology combinations were possible: 1) Emergency Abnormal & Radiology Normal; 2) Emergency Normal & Radiology Abnormal; 3) Both Abnormal; 4) Both Normal. For case 2) (where the patient had a “abnormal” classification from radiology and a “normal” diagnosis from ER) the patient was flagged as a possible missed diagnosis case. The breakdown of these four combinations for each hospital dataset is shown in Figure 2. The majority of patients had no abnormality recorded in both radiology and ER (light green, Both Normal), followed by patients with abnormalities recorded in both radiology and ER (green, Both Abnormal). A small number of cases were found with no abnormality in radiology but an abnormality in ER (orange, ER Abnormal – Radiology Normal); this occurred when the ER clinician suspected a condition but this turned out to be negative from the radiological assessment (and was, therefore, not an area of major concern). Finally, the number of patients flagged (red), out of the total number of patients, were: GCH 16 / 400 (4%), RCH 26 / 498 (5%) and RBWH 232 / 1480 (16%). The number of flagged cases was considerably higher than previous studies on quantifying missed fracture rates\(^1\) (especially for the RBWH dataset). To understand the reason behind this, we performed a manual analysis of all ER Normal – Radiology Abnormal (red) cases, reviewing both the radiology report, discharge diagnosis ICD-10 code and any associated ER notes. (The judgements were primary provided by the clinical author, KC.)

Our manual analysis showed that the ER discharge diagnosis ICD-10 code was often ambiguous — it may have
Table 5: Categories used in the manual review of all flagged cases (i.e., Emergency Normal & Radiology Abnormal).

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
<th>Example Diagnosis</th>
<th>Comment/Action Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrelated Diagnosis</td>
<td>The patient was assigned an ICD-10 diagnosis unrelated to a possible abnormality.</td>
<td>Gout, Self-harm, Congestive heart failure</td>
<td>Case requiring review by ER clinician.</td>
</tr>
<tr>
<td>Related Diagnosis</td>
<td>A condition was recorded that could relate to / cover an abnormality but it was not certain.</td>
<td>Crush injury, Laceration, Strain or sprain Fracture ISA Injury of the lower foot.</td>
<td>Case requiring review by ER clinician.</td>
</tr>
<tr>
<td>More General Diagnosis</td>
<td>More general condition that could cover an abnormality was coded, often as a result of imprecise ICD-10 coding.</td>
<td>Fracture ISA Injury of the lower foot.</td>
<td>Coding issue, currently requiring review by ER clinician but should address coding issue in the long term.</td>
</tr>
<tr>
<td>Missed Diagnosis</td>
<td>Real case of where ED clinician may have missed a limb abnormality.</td>
<td>No injury found, Patient did not wait.</td>
<td>Cases for actual follow up by ER clinician.</td>
</tr>
</tbody>
</table>

covered a fracture diagnosis (and thus an abnormality) but this was not explicit. For example, a patient with a fracture noted in their radiology report was discharged with S57 (Crushing injury of forearm). For this case, the clinician’s judgement was that it was not certain whether the crush injury actually indicated a fracture and therefore whether the ER clinician actually knew a fracture was present. S57 could not be added to the abnormal list as crush injury does not imply a fracture or other limb abnormality as defined here; indeed, there were many patients diagnosed with this code who did not have a fracture finding. Another common case was a discharge diagnoses of “strain or sprain” where an uncertain or very minor fracture was indicated in the radiology reports. For such cases, the ER clinician may have treated the patient (and therefore recorded the diagnosis) as having only a sprain/strain because the fracture was too minor or uncertain. Other flagged cases had a discharge diagnosis completely unrelated to limb injuries; e.g., an ICD-10 code representing self harm or congestive heart failure. To better understand the different flagged cases, all 274 were manually reviewed and assigned one of four categories described in Table 5. Note that all of the categories represent the situation where a limb abnormality (as defined in this work) may have been undiagnosed in ER; however, only the Missed Diagnosis category represents the situation where a limb abnormality was certainly missed.

The distribution of these different flagged categories, for each dataset, is shown in Figure 3. The most common category was a Related Diagnosis (e.g., the crush injury example). (All of the GCH and most of the RCH patients fell into this category.) The RBWH dataset contained a larger portion of both Unrelated Diagnosis and More General Diagnosis. This highlights the differences in the way the ICD-10 codes are assigned at different hospitals and how this might affect the use of these codes (our study being but one example of this). A total of 9 genuine Missed Diagnosis cases were identified in the RBWH dataset; these were either No injury found or Patient did not wait.

Clearly the way ICD-10 codes are assigned affects the reconciliation process, with many flagged cases being the situation where the ER clinician was aware of the abnormality but this fact was not conveyed in the ICD-10 code. However, even given this issue, if the clinician only had to review the 274 flagged cases, this would represent only 11% of the 2,378 reports they previously had to be reviewed — still representing a significant time saving.

Conclusions

We described a set of techniques to identify limb abnormalities from free-text radiology reports. The empirical evaluation showed that these methods are highly effective and that, importantly, they are robust across hospital datasets that are different in both size and similarity. We further show that the automatic classification can be used to reconcile the radiologist’s finding with the ICD-10 discharge diagnosis from the Emergency Room. Using this method, a number of potentially undiagnosed limb abnormalities were identified. A thorough manual analysis of these cases showed that some may be cases where the ICD-10 discharge diagnosis was ambiguous (highlighting the need for accurate ICD-10 coding in ERs); however, some genuine missed diagnoses were uncovered by the automated reconciliation process. Overall, the savings for a clinician were significant with only 11% of the entire dataset now requiring manual review.
Figure 3: Breakdown of different “flagged” cases (radiology abnormal but ER discharge diagnosis normal) according to the four different categories outline in Table 5. (Left plot show the absolute number of cases; the right plot shows the normalised portion of flagged cases.)

As such, the system is part of a pilot study in the Emergency Room of the Royal Brisbane and Women’s Hospital. While the final discharge diagnosis is recorded as an ICD-10 code, the ER clinician may also provide a short description. Although the ICD-10 code may be ambiguous (e.g., crush injury), the clinician’s description can contain an explicit mention of an abnormality. This additional source of information, in combination with the ICD-10 code, could be exploited to provide a better classification of the discharge diagnosis. In fact, similar machine learning techniques to those we have described for classifying free-text radiology reports could be adapted to classifying ER notes\textsuperscript{12}. The development and evaluation of such a method is an immediate area of future work.

Finally, this study has focused specifically on limb abnormalities described in radiology reports; however, the methods are not specific to this situation. Other types of abnormalities (e.g., presence of cancers) are currently being investigated and the methods are also being applied to the detection and reconciliation of different conditions mentioned in pathology reports (e.g., reconciling the antibiotic given to a patient against the antibiotic sensitivities identified in a microbiology report).

Acknowledgement

This research was supported by the Queensland Emergency Medicine Research Foundation Grant, EMPJ-11-158-Chu-Radiology.

References


Regex Rules for Keyword Spotting

List of regular expressions used to implement the keyword baseline method. In addition, if a match is found then the matching text is also tested for negation (e.g., “no fracture”), in which case the report is reported as “normal”, i.e., no fracture or other abnormality found.

"\bf\textbf{fracture}\textbf{\b}', "\bf\textbf{no}\textbf{\b}', "\bf\textbf{follow}[^\s]+\bf\b', "\bfx[^\s]+\bf\b', "\b\textbf{dislocation}\b', "\bf\b', "\b\textbf{osteomyelitis}\b', "\b\textbf{osteoly}\b', "\b\textbf{displacement}\b', "\b\textbf{intraarticular extension}\b', "\b\textbf{foreign body}\b', "\b\textbf{particular effusion}\b', "\b\textbf{avulsion}\b', "\b\textbf{septic arthritis}\b', "\b\textbf{sbuslukation}\b', "\b\textbf{osteotomy}\b', "\b\textbf{callus}\b', "\bf\b[a-z\s]+\b\bf\textbf{fracture}\b'"

ICD-10 Abnormal Codes

Set of ICD-10 codes that indicate an abnormal discharge diagnosis from the emergency room.

S03.2, S13.10, S03.0, S93.0, S93.30, S73.00, S83.10, S83.0, S93.10, S33.2, S33.10, S23.10, S43.1, S53.10, S63.10, S53.18, S43.3, S43.2, S63.0, S02.1, S62.1, S42.00, S02.9, S02.4, S02.6, S02.2, S02.3, S02.5, S72.40, S72.00, S72.10, S72.3, S72.04, S92.0, S92.9, S92.2, S02.0, S42.40, S42.3, S42.20, S62.2, S12.9, S12.8, S82.0, S32.4, S32.5, S32.83, S52.50, S52.8, S52.4, S62.0, S42.10, S62.5, S82.81, S82.6, S82.5, S82.88, S82.28, S82.82, S82.18, S22.5, S32.00, S22.3, S22.2, S22.00, S52.20, S82.4, S82.3, S62.6, S92.4, S92.5, M84.49, M86.99, M91.1, M93.0, Z47.8
Interpretable Probabilistic Latent Variable Models for Automatic Annotation of Clinical Text

Alexander Kotov, Ph.D.1, Mehedi Hasan, B.S.1, April Carcone, Ph.D2, Ming Dong, Ph.D.1, Sylvie Naar-King, Ph.D2, Kathryn Brogan-Hartlieb, Ph.D R.D.3

1Department of Computer Science, 2Pediatric Prevention Research Center, Wayne State University; 3Department of Dietetics and Nutrition, Florida International University

Abstract
We propose Latent Class Allocation (LCA) and Discriminative Labeled Latent Dirichlet Allocation (DL-LDA), two novel interpretable probabilistic latent variable models for automatic annotation of clinical text. Both models separate the terms that are highly characteristic of textual fragments annotated with a given set of labels from other non-discriminative terms, but rely on generative processes with different structure of latent variables. LCA directly learns class-specific multinomials, while DL-LDA breaks them down into topics (clusters of semantically related words). Extensive experimental evaluation indicates that the proposed models outperform Naïve Bayes, a standard probabilistic classifier, and Labeled LDA, a state-of-the-art topic model for labeled corpora, on the task of automatic annotation of transcripts of motivational interviews, while the output of the proposed models can be easily interpreted by clinical practitioners.

1. Introduction
Annotation or assignment of codes (labels) from a predefined codebook to fragments (entire documents or their parts) of clinical text is an integral part of medical practice and qualitative research. Such codes can be viewed as semantic labels, or high-level summaries (abstractions) of the raw textual data. Besides cataloging, such abstractions can facilitate the analysis of clinical text in general and clinical interview transcripts in particular. In this work, we focus on the transcripts of motivational interviews with obese adolescents conducted at a Pediatric Prevention Research Center (PPRC).

Childhood and adolescent obesity is a serious public health problem. Recent national data1 indicate that one in four children aged 2-5 years are overweight or obese. The trend of childhood obesity continues into adolescence – as of 2012, 18% of all adolescents and 23.7% of African American adolescents are obese1. Adolescents who are obese are likely to be obese as adults and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis2. Therefore, childhood and adolescence are critical periods for healthy eating and physical activity interventions to establish healthy weight gain trajectories. To design such interventions, PPRC clinicians conduct interviews with children and their caregivers grounded in the principles of Motivational Interviewing (MI)3, an evidence-based communication technique to increase intrinsic motivation and self-efficacy for behavior change. Detailed analysis of those interviews aims at identifying clinicians' communication strategies that are effective in triggering patient's motivational statements for the behavioral changes that will ultimately lead to weight loss. Recent literature reviews of mechanisms of effect in MI4 have concluded that clients' motivational statements about their own desire, ability, reasons and need for or commitment to change (or “change talk”) consistently predict actual behavior change5, as long as 34 months later6. Strategies to elicit motivational statements are typically identified via retrospective analysis of past interview transcripts. Part of this analysis involves assignment of codes to patient replies during the interviews. Analyzing sequences of assigned codes allows clinicians to better understand the patient’s thought process during the course of the interviews, without having to wade through entire transcripts over and over again. Such understanding, in turn, leads to further specification of the mechanisms of effect for intervention models, which can then be used to refine theory and guide clinical practice7.

Annotation of interview transcripts has traditionally been performed manually by trained coders, which is a tedious and resource intensive process. Therefore, methods that can efficiently and accurately distinguish the nuances of patient-provider communication can have a tremendous positive impact on many areas of clinical practice and research. Inferring psychological state of the patients during clinical interviews using only lexical content of their transcriptions is a challenging task for several reasons. First, many important indicators of emotions such as gestures, facial expressions and intonations are lost during the transcription process. Second, some utterances from patients during the interview may be too short and lack sufficient context for accurate classification. Furthermore, patients come from a variety of social, cultural, and educational backgrounds and their language is therefore quite
different. This problem is exacerbated when the interviews are conducted with children and adolescents, since children, in general, tend to often use incomplete sentences and frequently change subjects.

Automating the annotation of clinical documents is one of the fundamental problems in medical informatics, which can have tremendous implications for clinical practice. It falls under a general class of classification problems, which are typically addressed using supervised machine learning methods (or classifiers). Given a set of pre-classified data samples (called the training set) represented as a feature vector, in which each feature is the value of a feature function calculated based on a data sample, these methods learn the mapping from feature vectors to their classifications. Once learned, such mapping can be applied to classify new, unlabeled data samples (called the testing set). Classification problems arise in many different domains, from analysis of scientific literature and online reviews to digital forensics and medical informatics. Performance of different classifiers (including Naïve Bayes) on most common text classification tasks has been examined in detail in previous work. However, biomedical context places additional restriction of interpretability on machine learning models, as they are not only required to make correct classification decisions, but to also allow humans to easily understand how they arrived at these decisions. Interpretability of classification models is particularly important for psychological studies, such as Motivational Interviewing, since each class needs to have distinct interpretation (i.e. clearly correspond to a certain communication or behavior type). Furthermore, annotation models for these studies often need to be manually corrected. While the effectiveness of non-interpretable classifiers leveraging external resources, such as concepts from the Unified Medical Language System (UMLS) or clusters derived from a large external corpus, has been previously studied, there is still a need for designing interpretable models for annotating clinical interviews for behavioral studies, which typically contain very limited, domain-specific terminology and thus render general purpose medical lexicons ineffective for this task.

In this work, we focus on the problem of designing an interpretable model for automatic annotation of utterances in clinical interview transcripts with fine-grained semantic class (such as behavior type) and propose two new latent variable probabilistic models, Latent Class Allocation and Discriminative Labeled Latent Dirichlet Allocation as effective solutions to this problem. Both methods model how human annotators approach classification using probabilistic generative process. In particular, during training, they learn to distinguish the vocabularies that are highly indicative of the given classes from the general and non-discriminative terms via probabilistic assignment of latent variables to each term in the training corpus. The learned vocabularies in the form of multinomial distributions (or language models) are used to probabilistically classify new textual fragments and are easily interpretable by clinical practitioners. Although all experiments in this paper were conducted using clinical conversation data, the proposed methods can be applied to annotate any other type of clinical text.

2. Methods

2.1 Classes

As a golden standard for all experiments in this work, we used a sample of obesity Motivational Interview transcripts, in which patient utterances were manually annotated by human coders according to the "Minority Youth Sequential Code for Observing Process Exchanges" (MY-SCOPE) coding manual. Each utterance in the golden standard is labeled with one class. Among others, MY-SCOPE defines the following five classes of patient utterances, which correspond to major target patient behaviors clinicians were focused on when conducting obesity Motivational Interviews:

- **CL-**: negative commitment language;
- **CL+**: positive commitment language;
- **CT-**: negative change talk;
- **CT+**: positive change talk;
- **AMB**: ambivalence.

Commitment language (CL) is statements about patients’ intentions or plans for enacting weight related changes, which are positive, when supportive of behavior change, and negative, when against behavior change. Change talk (CT) corresponds to utterances that describe patients’ own desires, abilities, reasons, and need for adhering to weight loss recommendations and are also positive, when supportive of behavior change, and negative, when against behavior change. Ambivalent utterances (AMB) are change talk or commitment language statements that contain a
combination of positive and negative sentiments about changing one's behavior. Examples of utterances for each class are presented in Table 1.

Table 1. Examples of utterances representing the language and behavior types considered in this work.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL-</td>
<td>I eat a lot of junk food. Like, cake and cookies, stuff like that.</td>
</tr>
<tr>
<td>CL+</td>
<td>Well, I've been trying to lose weight, but it really never goes anywhere.</td>
</tr>
<tr>
<td>CT-</td>
<td>It can be anytime; I just don't feel like I want to eat (before) I'm just not hungry at all.</td>
</tr>
<tr>
<td>CT+</td>
<td>Hmm. I guess I need to lose some weight, but you know, it's not easy.</td>
</tr>
<tr>
<td>AMB</td>
<td>Fried foods are good. But it's not good for your health.</td>
</tr>
</tbody>
</table>

In the following sections, we present the classifiers used for the task of differentiating the above classes and report their performance in terms of standard evaluation metrics.

2.2 Features and baselines

We use standard bag-of-words feature generation framework, in which a predefined set of lexical features (or vocabulary), \( V = \{w_1, ..., w_N\} \), can appear in a given textual fragment. For example, one such feature could be the number of times a word (unigram) "exercise" appears in a given textual fragment. This way each textual fragment \( f \) is represented as a feature vector \( (n_{w_1,f}, ..., n_{w_N,f}) \), where \( n_{w,f} \) is the number of times feature (word) \( w \) occurs in \( f \).

To determine the best classification model for this task, in the following sections, we experimentally compare standard supervised machine learning methods, such as Naive Bayes\(^{13}\) and Labeled Latent Dirichlet Allocation\(^{18}\), with our proposed probabilistic classification models.

2.2.1 Naive Bayes

Naive Bayes (NB) is a standard probabilistic classifier, which annotates a given textual fragment \( f = \{w_1, ..., w_{N_f}\} \), consisting of \( N_f \) words, with a class \( c^* \), such that \( c^* = \arg \max c p(c|f) \), where \( p(c|f) \) is estimated by applying the Bayes' rule as follows:

\[
p(c|f) = \frac{p(f|c)p(c)}{p(f)} \propto p(f|c)p(c)
\]

In order to estimate \( p(f|c) \), Naive Bayes classifier makes an assumption about conditional independence of features given \( c \), the class of fragment \( f \):

\[
p(f|c) = \prod_{i=1}^{N_f} p(w_i|c)^{n_{w_i,f}}
\]

Despite its relative simplicity, NB has been experimentally demonstrated to be one of the most effective text classification algorithms ever created. In this work, we used a standard implementation of Multinomial NB algorithm from the Weka text mining toolkit.

2.3 Probabilistic models

We propose Latent Class Allocation (LCA) and Discriminative Labeled Latent Dirichlet Allocation (DL-LDA), two novel probabilistic generative latent variable models for the task of automatic coding of clinical interview transcripts, and compare their performance with Naive Bayes and Labeled Latent Dirichlet Allocation (L-LDA), a state-of-the-art probabilistic model for labeled data, on the task of annotating utterances in clinical text. LCA associates only one latent variable \( m \) with each word, which determines its type (whether a word is general or characteristic of a certain class). DL-LDA is an extension of L-LDA that makes a different set of assumptions about the structure of latent variables. Rather than directly associating each word with a latent variable determining its topic for a certain class, DL-LDA, similar to LCA, first associates with each word a latent variable, which determines whether a word is general or characteristic of a certain class and, only in the second case, associates it with another latent variable \( z \), determining its class-specific topic. Thus, DL-LDA can be viewed as a more structured version of Labeled LDA.
2.3.1 LCA

LCA models each textual fragment \( f \) labeled with class \( c_f \) as a set of alternating draws from a background multinomial \( \phi^{bg} \) that is drawn from a symmetric Dirichlet prior \( \beta^{bg} \) and a multinomial \( \phi^{cls} \) specific to \( c_f \) that is drawn from a symmetric Dirichlet prior \( \beta^{cls} \). The proportion of words drawn from \( \phi^{bg} \) and \( \phi^{cls} \) is controlled by a binomial distribution \( \lambda_f \). LCA generates annotated textual fragments according to the following probabilistic process:

1. draw \( \lambda_f \sim \text{Beta}(\gamma) \), a binomial distribution controlling the mixture of words in \( f \) drawn from the background and class-specific multinomials
2. for each word position \( i \) of \( N_f \) in \( f \):
   a) draw Bernoulli switching variable \( m_{ki} \sim \lambda_f 
   b) if \( m_{ki} = bg \):
      - draw a word \( w_{ki} \sim \phi^{bg} 
   c) if \( m_{ki} = cls \):
      - draw a word \( w_{ki} \sim \phi^{cls,c_f} 

The generative process of LCA in plate notation is presented in Figure 1a.

![LCA Diagram](image)

**Figure 1. Generative processes of the proposed and baseline latent variable models in plate notation.**

Annotation of textual fragments in the testing set with LCA is done using class-specific multinomials \( \phi^{cls} \) (or \( p(w|c) \)) determined as a result of posterior inference on the training set to derive \( p(c|w) \), distributions showing how indicative each word \( w \) is for each class \( c \):

\[
p(c|w) = \frac{p(w|c)p(c)}{p(w)}
\]

where \( p(c) = \frac{n_{fc}}{M} \) (\( n_{fc} \) is the number of interview fragments labeled with class \( c \) and \( M \) is the total number of fragments) and \( p(w) \) is a probability of word \( w \) in a collection language model estimated using maximum likelihood.

\( p(c|w) \) are then used to classify \( f \) according to the following formula:

\[
c^* = \arg \max_c p(c|f) = \prod_{i=1}^{N_f} p(c|w_i)^{n_{wi,f}}
\]

2.3.3 L-LDA

L-LDA directly associates a latent variable \( z \) with each word that determines its assignment to a topic specific to \( c_f \). It is state-of-the-art topic model for labeled textual collections that has been shown to outperform standard classifiers, such as SVM, for the task of multi-class classification\(^{18}\). The generative process of L-LDA in plate notation is presented in Figure 1b. L-LDA along with NB is used as a baseline in our experimental evaluation.
2.3.2 DL-LDA

DL-LDA models each textual fragment \( f \) labeled with class \( c_f \) as a mixture of the background topic \( \phi^{bg} \) drawn from a symmetric Dirichlet prior \( \beta^{bg} \) and \( K^{cls} \) topics drawn from a uniform Dirichlet prior \( \beta^{cls} \). DL-LDA generates the textual fragments in clinical interviews according to the following probabilistic process:

1. draw \( \lambda_f \sim \text{Beta}(\gamma) \), a binomial distribution controlling the mixture of background and class-specific topics for \( f \)
2. draw \( \Theta_f^{cls} \sim \text{Dir}(\alpha^{cls}) \), a distribution of class-specific topics for \( f \)
3. for each word position \( i \) of \( N_f \) in \( f \):
   (a) draw Bernoulli switching variable \( m_{ki} \sim \lambda_f 
   (b) if \( m_{ki} = \text{bg} \):
      - draw a word \( w_{ki} \sim \phi^{bg} 
   (c) if \( m_{ki} = \text{cls} \):
      - draw a topic \( z_{ki} \sim \Theta_f^{cls} 
      - draw a word \( w_{ki} \sim \phi^{cls, cf} 

The generative process of DL-LDA in plate notation is presented in Figure 1c. Classification using DL-LDA is performed by first deriving a class-specific multinomial \( p(w|c) \) per each class \( c \) from class-specific topics \( \phi^{cls, cf} \) (or \( p(w|c, z) \)) by marginalizing over \( z \):

\[
p(w|c) = \sum_{z=1}^{K^{cls}} p(w|c_i, z)
\]

and then using class-specific multinomials to directly classify \( f \), similar to LCA.

We would like to note that the inference algorithm for LCA, L-LDA and DL-LDA is adaptable to distributed environment and therefore the proposed methods can be scaled up to very large datasets\(^{19,20}\).

3. Results

The dataset used for experiments in this work consists of 2966 manually annotated fragments of interview transcripts. The distribution of the number of samples per each class is shown in Table 2.

<table>
<thead>
<tr>
<th>Class</th>
<th># Samples</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL-</td>
<td>73</td>
<td>2.46 %</td>
</tr>
<tr>
<td>CL+</td>
<td>875</td>
<td>29.50 %</td>
</tr>
<tr>
<td>CT-</td>
<td>278</td>
<td>9.37 %</td>
</tr>
<tr>
<td>CT+</td>
<td>1657</td>
<td>55.87 %</td>
</tr>
<tr>
<td>AMB</td>
<td>83</td>
<td>2.80 %</td>
</tr>
<tr>
<td>Total</td>
<td><strong>2966</strong></td>
<td><strong>100 %</strong></td>
</tr>
</tbody>
</table>

The dataset was first pre-processed by removing very frequently occurring terms (that occur in more than 25% of textual fragments). We also used the following pre-processing methods to study their effect on classification performance:

- **RAW**: no preprocessing, original dataset is used;
- **STEM**: Porter stemmer is applied to each term in the dataset to eliminate morphological variation;
- **STOP**: stopwords are removed, but stemming is not applied;
- **STOP-STEM**: Porter stemmer is applied to each term in the dataset and stopwords are removed.

For all experiments we used randomized 5-fold cross-validation. The Gibbs sampler for posterior inference of parameters of LCA, L-LDA and DL-LDA was run for 1000 iterations. Classification performance of NB, L-LDA, LCA and DL-LDA on the task of differentiating 5 language categories when experimental dataset is pre-processed with different methods is summarized in Tables 3, 4, 5 and 6, respectively.

Table 3. Performance of Naïve Bayes using different pre-processing methods. Best result for each performance metric is highlighted in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAW</td>
<td>0.522</td>
<td>0.523</td>
<td>0.506</td>
</tr>
<tr>
<td>STEM</td>
<td>0.534</td>
<td>0.534</td>
<td>0.518</td>
</tr>
<tr>
<td>STOP</td>
<td>0.511</td>
<td>0.526</td>
<td>0.510</td>
</tr>
<tr>
<td>STOP-STEM</td>
<td>0.510</td>
<td>0.519</td>
<td>0.506</td>
</tr>
</tbody>
</table>

Table 4. Performance of Labeled Latent Dirichlet Allocation using different pre-processing methods. Best result for each performance metric is highlighted in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAW</td>
<td>0.537</td>
<td>0.530</td>
<td>0.480</td>
</tr>
<tr>
<td>STEM</td>
<td>0.544</td>
<td>0.540</td>
<td>0.474</td>
</tr>
<tr>
<td>STOP</td>
<td>0.530</td>
<td>0.520</td>
<td>0.478</td>
</tr>
<tr>
<td>STOP-STEM</td>
<td>0.538</td>
<td>0.517</td>
<td>0.475</td>
</tr>
</tbody>
</table>

Table 5. Performance of Latent Class Allocation using different pre-processing methods. Best result for each performance metric is highlighted in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAW</td>
<td>0.543</td>
<td>0.534</td>
<td>0.537</td>
</tr>
<tr>
<td>STEM</td>
<td>0.557</td>
<td>0.542</td>
<td>0.549</td>
</tr>
<tr>
<td>STOP</td>
<td>0.541</td>
<td>0.508</td>
<td>0.520</td>
</tr>
<tr>
<td>STOP-STEM</td>
<td>0.543</td>
<td>0.515</td>
<td>0.525</td>
</tr>
</tbody>
</table>

Table 6. Performance of Discriminative Labeled Latent Dirichlet Allocation using different pre-processing methods. Best result for each performance metric is highlighted in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAW</td>
<td>0.591</td>
<td>0.533</td>
<td>0.537</td>
</tr>
<tr>
<td>STEM</td>
<td>0.586</td>
<td>0.515</td>
<td>0.527</td>
</tr>
<tr>
<td>STOP</td>
<td>0.560</td>
<td>0.504</td>
<td>0.508</td>
</tr>
<tr>
<td>STOP-STEM</td>
<td>0.557</td>
<td>0.492</td>
<td>0.498</td>
</tr>
</tbody>
</table>

Table 7. Summary of the best performance of different methods for the task of annotation of 5 original language types. Best result for each performance metric is highlighted in boldface.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.522</td>
<td>0.523</td>
<td>0.506</td>
</tr>
<tr>
<td>LCA</td>
<td>0.543</td>
<td>0.534</td>
<td>0.537</td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.537</td>
<td>0.530</td>
<td>0.480</td>
</tr>
<tr>
<td>DL-LDA</td>
<td>0.591</td>
<td>0.533</td>
<td>0.537</td>
</tr>
</tbody>
</table>
Since classification accuracy of DL-LDA is dependent on the number of per-class topics, which is a parameter that needs to be specified a priori, we first optimized it with respect to F1 score. Figure 2 indicates that the optimal classification results for DL-LDA in combination with different pre-processing methods are achieved when the number of topics is small (2 or 3 in most cases).

![Figure 2. F1 score of DL-LDA by varying the number of topics and in combination with different pre-processing methods on the task of classification of all language categories.](image)

The best results for each proposed model and the baselines are summarized and compared in Table 7, while the per-class breakdown of the best results for all models is provided in Table 8.

**Table 8. Summary of the best per class performance in terms of F1 score of different classifiers for the task of distinguishing 5 original language types. Best result for each class is highlighted in boldface.**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td><strong>0.129</strong></td>
<td>0.329</td>
<td>0.164</td>
<td>0.691</td>
<td>0.170</td>
</tr>
<tr>
<td>LCA</td>
<td>0.094</td>
<td><strong>0.437</strong></td>
<td><strong>0.223</strong></td>
<td>0.682</td>
<td><strong>0.162</strong></td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.025</td>
<td>0.252</td>
<td>0.066</td>
<td>0.708</td>
<td>0.128</td>
</tr>
<tr>
<td>DL-LDA</td>
<td>0.025</td>
<td>0.396</td>
<td>0.114</td>
<td><strong>0.729</strong></td>
<td>0.061</td>
</tr>
</tbody>
</table>

In the second set of experiments, we aggregated the interview fragments labeled as positive and negative commitment language (CL+ and CL-) and change talk (CT+ and CT-) into one combined class for commitment language (CL) and one combined class for change talk (CT), respectively, and evaluated the accuracy of our classifiers in distinguishing the interview fragments labeled with the resulting three broader classes.

**Table 9. Number of samples per aggregated positive and negative sub-classes of CL and CT.**

<table>
<thead>
<tr>
<th>Class</th>
<th>Samples</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL</td>
<td>948</td>
<td>31.96 %</td>
</tr>
<tr>
<td>CT</td>
<td>1935</td>
<td>65.24 %</td>
</tr>
<tr>
<td>AMB</td>
<td>83</td>
<td>2.80 %</td>
</tr>
<tr>
<td>Total</td>
<td><strong>2966</strong></td>
<td><strong>100 %</strong></td>
</tr>
</tbody>
</table>

Distribution of samples across these three classes is shown in Table 9. For this task we used raw data for each classifier (no preprocessing). We optimized the number of topics for DL-LDA with respect to the F1 score (Figure 3) and found out that again the optimal number of topics is 3. Performance of different classifiers on the task of differentiating the interview fragments labeled with CL, CT and AMB is summarized in Table 10. In the third set of experiments, we aggregated the interview fragments labeled as positive sub-classes of commitment language (CL+) and change talk (CT+) into one combined positive class (+) and negative sub-classes of commitment language (CL-) and change talk (CT-) into one combined negative class (-) and evaluated the accuracy of our classifiers in distinguishing the interview fragments labeled with the resulting sentiment modality-based broader classes.
Figure 3. F1 score of DL-LDA by varying the number of topics on the task of classification of aggregated positive and negative sub-classes within CL and CT.

Table 10. Performance of the proposed methods and the baselines on the task of distinguishing aggregated positive and negative sub-classes within CL and CT. Best result for each metric is highlighted in boldface.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.617</td>
<td>0.627</td>
<td>0.611</td>
</tr>
<tr>
<td>LCA</td>
<td>0.674</td>
<td>0.651</td>
<td>0.656</td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.634</td>
<td>0.631</td>
<td>0.587</td>
</tr>
<tr>
<td>DL-LDA</td>
<td>0.673</td>
<td>0.637</td>
<td>0.633</td>
</tr>
</tbody>
</table>

Distribution of samples across these three classes in shown in Table 11. Similarly to the task of differentiating CT, CL and AMB, we tuned DL-LDA with respect to F1 score (Figure 5) and found out that this time the optimal number of per-class topics is 5.

Figure 5. F1 score of DL-LDA by varying the number of topics on the task of classification of aggregated positive and negative sub-classes across CL and CT.

Performance of different classifiers on the task of differentiating the utterances labeled with +, - and AMB is summarized in Table 12.

4. Discussion

Several important observations can be made from Tables 3, 4, 5 and 6. First, stemming and stopwords removal
degrade classification performance of DL-LDA, which performs best without any pre-processing, while Naïve Bayes, LCA and L-LDA achieve the best classification performance when stemming is applied.

Table 11. Number of samples per aggregated positive and negative sub-classes across CL and CT.

<table>
<thead>
<tr>
<th>Class</th>
<th># Samples</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>351</td>
<td>11.83 %</td>
</tr>
<tr>
<td>+</td>
<td>2532</td>
<td>85.37 %</td>
</tr>
<tr>
<td>AMB</td>
<td>83</td>
<td>2.80 %</td>
</tr>
<tr>
<td>Total</td>
<td>2966</td>
<td>100 %</td>
</tr>
</tbody>
</table>

Table 12. Performance of the proposed methods and the baselines on the task of distinguishing aggregated positive and negative sub-classes across CL and CT. Best result for each metric is highlighted in boldface.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.734</td>
<td>0.778</td>
<td>0.753</td>
</tr>
<tr>
<td>LCA</td>
<td>0.818</td>
<td>0.771</td>
<td>0.790</td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.814</td>
<td>0.774</td>
<td>0.781</td>
</tr>
<tr>
<td>DL-LDA</td>
<td>0.838</td>
<td>0.770</td>
<td>0.793</td>
</tr>
</tbody>
</table>

However, for all 4 classifiers used in this work stopwords removal by itself and in combination with stemming decreases the accuracy of classification, which suggests that common stopwords might be important indicators for some of the language categories. Second, as follows from Table 7, LCA and DL-LDA outperform both baselines (NB and L-LDA) in terms of all three performance measures (Recall, Precision and F1 score). Across all models, DL-LDA achieves the best performance in terms of Recall, while LCA achieves the best performance in terms of both Precision and F1 score. As follows from Table 8, LCA and DL-LDA also achieve the best per-class performance for 4 out of 5 classes. These results lead to two important conclusions. First, explicitly accounting for discriminativity of terms (general or class-specific) in an utterance with a latent variable allows to improve annotation performance using probabilistic methods. Second, additional division of class-specific multinomials into class-specific topics by DL-LDA allows to improve recall, but not precision and F1 score.

Results of classifying language types without taking into account modality (Table 10) indicate that LCA is particularly suited for this task and again support our assumption about the utility of differentiating the terms by their discriminativity. LCA and DL-LDA use the strength of statistical associations of terms with the class labels as a measure of their discriminativity. Since non-discriminative words are the ones that occur in many fragments labeled with different classes, statistical associations of these terms with class labels are relatively weak, which is effectively captured by LCA and DL-LDA.

Table 13. Most characteristic words for each utterance label according to LCA.

<table>
<thead>
<tr>
<th>Class</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL-</td>
<td>drink sugar gatorade lot hungry splenda beef tired watch tv steroids sleep home nervous confused starving appetite asleep craving pop fries computer</td>
</tr>
<tr>
<td>CL+</td>
<td>stop run love tackle vegetables efforts juice swim play walk salad fruit</td>
</tr>
<tr>
<td>CT-</td>
<td>got laughs sleep wait answer never tired splenda fault phone joke weird hard don’t</td>
</tr>
<tr>
<td>CT+</td>
<td>time go mom brother want happy clock boy can move library need adopted reduce sorry solve overcoming lose</td>
</tr>
<tr>
<td>AMB</td>
<td>what taco mmm know say plus snow pain weather</td>
</tr>
</tbody>
</table>

Results of classifying modality (Table 12) indicate that DL-LDA is the best in detecting the attitude of the speaker. This can be explained by the fact that only a portion of vocabularies indicative of specific classes reflect the sentiment modality of an utterance, therefore splitting class-specific multinomials into class-specific topics by DL-LDA allows to isolate the sentiment-specific vocabularies and leverage them during classification.
Examples of the most characteristic terms for each utterance label determined by LCA are provided in Table 13. As follows from Table 13, negative commitment language is strongly associated with the words reflecting poor diet ("sugar", "pop", "fries") and sedentary lifestyle ("watch", "tv", "computer"), while positive commitment language is strongly associated with the terms related to exercise ("walk", "play", "run") and healthy food options ("salad", "vegetables", "fruit"). The words characteristic of CT- and CT+ generally reflect negative ("don’t", "never", "tired") and positive ("can", "need", "lose", "happy") attitudes towards weight loss, respectively.

5. Conclusion

In this paper, we proposed Latent Class Allocation, a novel interpretable probabilistic model for supervised text classification, and Discriminative Labeled LDA, an extension of Labeled LDA, that differentiates between class-specific and general terms. Through extensive experimental evaluation, we demonstrated that the proposed models have consistently better performance for the task of single class annotation of fragments of Motivational Interviewing transcripts than state-of-the-art methods, such as Naive Bayes and Labeled LDA.

References

Representation of Functional Status Concepts from Clinical Documents and Social Media Sources by Standard Terminologies

Jinqiu Kuang, M.S. 1, April F. Mohanty, Ph.D. 2, M.P.H., Rashmi V. H., M.Tech. 1, Charlene R. Weir, Ph.D., R.N. 1,2, Bruce E. Bray, M.D. 1, Qing Zeng-Treitler, Ph.D. 1,2

1Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, UT, USA
2Informatics Decision-Enhancement and Analytic Sciences (IDEAS) Center, George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, UT, USA

Abstract

Patient-reported functional status is widely recognized as an important patient-centered outcome that adds value to medical care, research, and quality improvement. Functional status outcomes are, however, not routinely or uniformly collected in the medical record, except in certain small patient populations (e.g., geriatrics, nursing home residents). To utilize patient reported functional status for clinical research and practice, we manually collected 2,763 terms from clinical records and social media sites and modeled them on the widely used Short Form-36 Health Survey. We then examined the coverage of the Unified Medical Language System (UMLS) for these functional status terms through automated mapping. Most terms (85.9%) did not have exact matches in the UMLS. The partial matches were prevalent, however, they typically did not capture the terms’ exact semantics. Our study suggests that there is a need to extend existing standard terminologies to incorporate functional status terms used by patients and clinicians.

Introduction

Patient reported functional status is widely recognized as an important patient-centered outcome that adds value to medical care, research, and quality improvement. For instance, during the late 1980s researchers observed that patients’ functional status was correlated to their compliance with antihypertensive treatment, even when treatment led to successful blood pressure control. Reduction in functional status is frequently the first sign of declining health for patients with chronic conditions and is related to the severity of acute illness and intensity of resource use. Functional status is a prognostic predictor of future risk of hospitalization readmission, morbidity and mortality. Functional status information is central to many healthcare decisions, including end-of-life care, living arrangements, and patient-tailored treatment. Some experts assert that functional status should be included as the “sixth vital sign”.

In the context of atrial fibrillation and heart failure, alteration to functional status is one of the most commonly reported symptoms, including reductions in exercise tolerance of 15-20% among patients diagnosed with atrial fibrillation. Measurement of functional status can help to inform diagnosis, prognosis, and can guide care/treatment strategies related to these conditions, in addition to facilitating examinations of treatment response.

While instruments such as the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) have been developed to capture patient reported functional status, functional status is not routinely or uniformly collected in the medical record. At the same time, clinical records as well as social media data do contain a large number of descriptions of patient functional status.

There has been much interest in and effort toward development of methods to extract functional status terms, using existing frameworks, such as the International Classification of Functioning, Disability and Health (ICF), especially for healthcare delivery for nursing home and rehabilitation patients. The ICF was introduced 2001 by the World Health Organization as “a unified and standard language and framework for the description of health and health-related states”. This standard terminology is currently included in the Unified Medical Language System (UMLS), a compendium of over 100 national and international vocabularies and classifications.

As these and other standard terminologies have been developed, it is important to update findings related to their coverage of functional status concepts. Also, since prior work on functional status concepts has been limited to the rheumatology care setting, there is a need to extend these findings to other settings including other medical care settings and consumer health.
In this study, we specifically examined functional status documented in the electronic medical record (EMR) and social media for patients with cardiovascular diagnoses such as atrial fibrillation. The terms were categorized and mapped to standard clinical terminologies. We also analyzed the causes of non-matches and partial matches.

Methods

Data Collection

We defined functional status terms as a span of text that expresses an idea or description corresponding to functional status. To collect functional status terms, we used clinical records and social media, both of which have been increasingly utilized by biomedical researchers. In social media, patients voluntarily describe functional status in their own terms. In the EMR, patient reports are often paraphrased and sometimes clinicians document functional status using a variety of instruments.

EMR data were obtained from the Veterans Administration Informatics and Computing Infrastructure (VINCI) database\(^1\). The Institutional Review Board approved the use of electronic medical data from VINCI and no human subjects were contacted for this study. We randomly selected a set of 800 clinical documents of patients who were diagnosed with atrial fibrillation or atrial flutter for review. Clinical documents were free text documents or progress notes such as cardiology visit notes, primary care nursing/physician notes, telephone encounter notes, palliative/geriatric medicine notes, and discharge summaries. Generally, these clinical documents do not use standardized terminologies to convey information regarding functional status.

Social media data were obtained from three different online discussion forums for cardiovascular diseases and atrial fibrillation. These sources were: (1) Atrial Fibrillation Support Group on dailystrength.org, (2) Heart & Cardiovascular Disease forum on healingwell.com, and (3) Cardiovascular Disease Prevention Expert Forum on medhelp.org. Questions, replies and comments posted by patients were extracted from the discussion forums using Web Scraper, a Chrome browser extension used for extraction of data from web pages. From the extracted content, 150 latest posts from source 1 and 100 latest posts each from source 2 and source 3 were chosen.

For both datasets, one investigator then manually reviewed clinical documents or posts containing word/phrases that indicated functional status of the patient according to the SF-36 Health Survey, with the intent to identify all potential candidate terms. Then a second investigator reviewed all the extracted terms. When a consensus could not be reached, a third adjudicator was included in the extraction of functional status terms.

We excluded duplicate words/phrases and lists of unique functional status terms from VINCI and social media data sources were created. For the VINCI data, we collected 1,050 functional status terms and removed duplicate terms. A total of 974 unique terms were mapped to UMLS. For the social media data, a total of 1,623 functional status terms were extracted out of which 980 terms were unique.

Classification of Functional Status Terms Based on SF-36 Health Survey Scales

To understand the types functional status terms being used in medical records and social media, we categorized them according to the SF-36 Health Survey scales. The SF-36 Health Survey includes eight scales: Physical Functioning, Role-Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role-Emotional and Mental Health (http://www.sf-35.org/tools/SF36.shtml). Scores for each of these scales are determined by all but one (self-reported health transition) of the 36 items of the SF-36 Measurement Model. There are two summary measures of the SF-36 Health Survey that the eight scales aggregate to. The Physical Health summary measure includes Physical Functioning, Role-Physical, Bodily Pain, and General Health. The Mental Health summary measure includes Vitality, Social Functioning, Role-Emotional, and Mental Health.

Using the 35 items listed for the eight scales as a reference, three investigators jointly reviewed a subset of VINCI and social media terms to develop a classification guideline for assigning each term to one of the eight SF-36 Health Survey scales. The three investigators divided the remaining terms (two investigators reviewed the VINCI terms and one reviewed the social media terms) to independently review and assign them to one of the eight SF-36 Health Survey scales and based on the classification guideline.
Classification of Type of Match by Standard Terminologies

For the purpose of mapping functional status terms to the UMLS Metathesaurus, we used the MetaMap system developed by the National Library of Medicine (NLM)\(^\text{18}\). Unique functional status terms from VINC1 and social media were parsed using MetaMap to determine the corresponding concepts in UMLS. The parsed terms, mapped concepts (along with concept unique identifiers [CUI]) and respective UMLS source vocabularies were recorded. Three investigators divided the functional status terms to independently review and categorize them as one of three types of matches: a 'complete match', a 'partial match', or a 'no match'. Terms where the entire word/phrase mapped to a single concept in UMLS were categorized as 'complete match'. Compound terms consisting of two or more words where only a few words mapped to UMLS concepts were categorized as 'partial match'. For example a ‘partial match’ could include “jogging (6 mph)”, where “jogging” mapped to the UMLS but the qualifier “(6 mph)” did not map. Terms that did not map to any UMLS concepts were categorized as 'no match', an example of a term that did not map to any UMLS concepts was “debilitated”. We further assessed whether complete matches were correct or incorrect. We defined complete incorrect matches as terms where the entire word/phrase mapped to a single concept in UMLS and did not capture the correct semantic meaning. An example of a complete incorrect match would be the term “fine”, where the incorrect semantic meaning would be a financial fine versus feeling fine.

Review of Classifications

Once all of the functional status terms were manually assigned to one of the eight SF-36 Health Survey scales and were categorized as a complete-, partial-, or no-match, the three investigators exchanged lists of classified terms to reach a consensus on the final SF-36 Health Survey scale assignment and type of match for each term. Where there was a disagreement between two investigators, the final assignment was reached by including a third adjudicator.

Data Analysis

We examined the overlap of functional status terms and matched concepts across the two sources (i.e. VINC1 or social media). For each source of functional status terms we examined the frequency distributions of functional status terms across SF-36 Health Survey scales. Percentages of partially matched terms, completely matched terms and terms that yielded no match under each of the SF-36 Health Survey scale were determined. We also calculated the percent coverage of functional status concepts by various vocabulary sources of the UMLS Metathesaurus.

Results

Description of Functional Status Concepts in VHA and Social Media Sources

Overall, there was an overlap of 14 terms (exact term match) across both functional status term sources (i.e. VINC1 and social media data). The overlap between mapped UMLS concepts from VINC1 and social media data was much larger - 44.7% (or \(n = 378\) concepts) of the mapped concepts from social media data were also found in VINC1 data.

The most common SF-36 Health Survey scales that the extracted 974 VHA terms were classified to were Physical Functioning (37.9%), Role-Physical (23.9%), and Mental Health (16.3%), Figure 1a. The top three SF-36 Health Survey scales for the social media terms were Physical Functioning (31.9%), Mental Health (23.5%), and Role-Physical (13.5%), Figure 1b. Fewer than 2.0% of terms were classified to Role-Emotional or Social Functioning scales, regardless of source of functional status term.
Figure 1. The Distribution of Three Types of Match by the Eight SF-36 Health Survey Scales
Abbreviations: VINCI, Veterans Affairs Informatics and Computing Infrastructure; BP, Bodily Pain; GH, General Health; MH, Mental Health; PF, Physical Functioning; RE, Role-Emotional; SF, Social Functioning; VT, Vitality

Coverage of Functional Status Concepts by Standard Terminologies

The comparison of the number of functional status terms that fell into ‘complete match’, ‘no match’ and ‘partial match’ categories in VINCI and social media data is shown in Figure 1 and Table 1. Overall, 14.1% of functional status terms were completely matched and 83.2% were partially matched to UMLS. Social media data contained a larger number of ‘complete match’ and ‘no match’ terms when compared to VINCI data. Most complete matches were in the Mental Health scale SF-36 Health Survey regardless of source of functional status term (Figure 1). Most match failures (‘partial-‘ and ‘no-match’ combined) corresponded to the Physical Functioning SF-36 Health Survey scale, regardless of data source.

Table 1. Overall Match Percent for Terms from VINCI and Social Media Data Sources

<table>
<thead>
<tr>
<th>Match Type</th>
<th>VINCI N (%)</th>
<th>Social Media N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>123 (12.6)</td>
<td>153 (15.6)</td>
</tr>
<tr>
<td>No match</td>
<td>12 (1.2)</td>
<td>40 (4.1)</td>
</tr>
<tr>
<td>Partial</td>
<td>839 (86.1)</td>
<td>787 (80.3)</td>
</tr>
</tbody>
</table>

Abbreviations: VINCI, Veterans Affairs Informatics and Computing Infrastructure

The UMLS Metathesaurus is made up of more than 100 national and international source vocabularies and classifications. A concept can be found in one or multiple vocabularies. Mapped concepts from VINCI and social media data were spread across 43 different source vocabularies in UMLS. Table 2 lists the top 10 source vocabularies with the most mapped concepts. The top four vocabularies that had the most coverage were MTH, SNOMEDCT_US, National Cancer Institute (NCI) and Consumer Health Vocabulary (CHV) for both VINCI and social media data. Combination of these four vocabularies covered 96.2% of the mapped concepts from VINCI data and 97.3% of the mapped concepts from social media data.
Table 2. Top 10 UMLS Source Vocabularies for VINCI and Social Media Data Sources

<table>
<thead>
<tr>
<th>Source Vocabulary (%) Coverage</th>
<th>VINCI</th>
<th>Social Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTH (78.2)</td>
<td>MTH (81.9)</td>
<td></td>
</tr>
<tr>
<td>SNOMEDCT_US (48.6)</td>
<td>SNOMEDCT_US (51.9)</td>
<td></td>
</tr>
<tr>
<td>NCI (40.2)</td>
<td>NCI (47.6)</td>
<td></td>
</tr>
<tr>
<td>CHV (34.8)</td>
<td>CHV (36.4)</td>
<td></td>
</tr>
<tr>
<td>MSH (18.2)</td>
<td>MSH (18.3)</td>
<td></td>
</tr>
<tr>
<td>NLMSubSyn (17.0)</td>
<td>NLMSubSyn (15.2)</td>
<td></td>
</tr>
<tr>
<td>SNMI (6.8)</td>
<td>AOD (11.8)</td>
<td></td>
</tr>
<tr>
<td>LNC (6.2)</td>
<td>SNMI (10.7)</td>
<td></td>
</tr>
<tr>
<td>AOD (5.6)</td>
<td>LNC (4.6)</td>
<td></td>
</tr>
<tr>
<td>HL7V3.0 (3.1)</td>
<td>ICF (4.1)</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: VINCI, Veterans Affairs Informatics and Computing Infrastructure

Some functional status terms from VINCI and social media data yielded complete matches, most of which were correctly matched. There were 123 complete matched terms from VINCI out of which 120 were correctly mapped and were found in 26 UMLS source vocabularies. Social media data had 153 complete matches out of which 141 mapped to correct concepts. These concepts were found in 24 different vocabulary sources of the UMLS Metathesaurus. Percentages of coverage of completely matched correct concepts by 10 source vocabularies that had maximum coverage for VINCI data and social media data are reported in Table 2b. SNOMEDCT_US, MTH, CHV and MSH were the top 4 source vocabularies with respect to coverage of completely matched terms that were mapped to correct concepts for both VINCI data and social media data. Combination of these top 4 source vocabularies contained 96.7% and 93.61% of the correctly mapped concepts from VINCI data and social media data, respectively.

Table 3. Top 10 UMLS Source Vocabularies for Complete and Correct Matches

<table>
<thead>
<tr>
<th>Source Vocabulary (%) Coverage</th>
<th>VINCI</th>
<th>Social Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNOMEDCT_US (65.8)</td>
<td>SNOMEDCT_US (66.0)</td>
<td></td>
</tr>
<tr>
<td>MTH (63.3)</td>
<td>MTH (65.3)</td>
<td></td>
</tr>
<tr>
<td>CHV (50.8)</td>
<td>CHV (55.3)</td>
<td></td>
</tr>
<tr>
<td>MSH (41.7)</td>
<td>MSH (46.1)</td>
<td></td>
</tr>
<tr>
<td>NLMSubSyn (38.3)</td>
<td>NCI (35.5)</td>
<td></td>
</tr>
</tbody>
</table>
Most of the partial matches did not capture the exact semantics of the terms. We did not attempt to assess the correctness of the partial matches because frequently, a part that matched could be considered to be correct while another part could not. For instance, for the term “back to normal” the term “back” mapped to a number of concepts none of which were correct and the term “normal” mapped to the concept “Normal” which was correct. Together, they also failed to represent the complete semantic meaning.

We explored reasons for failures for a subset of non-matches and partial matches, these examples are presented in Table 4.

Table 4. Examples of Type of Match to UMLS and Reasons For Failures

<table>
<thead>
<tr>
<th>Type of Match</th>
<th>Examples</th>
<th>Reason for Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Match</td>
<td>Too zonked</td>
<td>Concept does not exist in UMLS</td>
</tr>
<tr>
<td></td>
<td>Debilitated</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Annoying</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cowering</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unfocused</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Go to gym</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doing OK</td>
<td>Use of colloquial term</td>
</tr>
<tr>
<td>Partial Match</td>
<td>Jogging (6 mph)</td>
<td>No match for qualifier/modifier</td>
</tr>
<tr>
<td></td>
<td>Extreme fatigue</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Have problems tinkering around in the garage and outside</td>
<td>Too many descriptive details</td>
</tr>
<tr>
<td></td>
<td>Has difficulty taking arm to the side and away from the body</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exercises daily on treadmill alternating with walking 3-4 miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Feels lousy</td>
<td>Use of colloquial term</td>
</tr>
<tr>
<td></td>
<td>Drives me nuts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PAIN BEST: 4/10</td>
<td>Use of templates (specific to VINCI data)</td>
</tr>
<tr>
<td></td>
<td>Total [8]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Feel like someone’s choking me</td>
<td>Use of metaphors (specific to social media data)</td>
</tr>
<tr>
<td></td>
<td>Feels like death taking over</td>
<td></td>
</tr>
<tr>
<td></td>
<td>In a fog</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: VINCI, Veterans Affairs Informatics and Computing Infrastructure

Discussion

With the goal of utilizing patient-reported functional status in research and clinical practice, we collected and analyzed 1,954 unique functional status terms extracted from VINCI EMR data among patients diagnosed with atrial fibrillation or atrial flutter and from social media posts of three cardiovascular disease-related online discussion forums. Overlap of the exact term across the two sources of functional status terms (i.e. EMR clinical documents containing free text data or social media posts) was small although, further examination of the overlap of mapped
UMLS functional status concepts across sources was much larger. The most common types of terms according the eight scales of the SF-36 Health Survey were Physical Functioning (34.9% of all terms) and Mental Health (20% of all terms). The least common types of terms were Role-Emotional (<0.1% of all terms) and Social Functioning (1.0% of all terms).

Overall, the coverage of functional status terms by current standard clinical terminologies was low: only 12.6% of terms extracted from VHA and 15.6% of terms extracted from social media were completely mapped to UMLS concept names. Some complete matches were not correct. Examples include the term “work out” (as in exercise) was mapped to “Unemployment” and the term “cardio” (again, as in exercise) was mapped to “Heart”. Partial matches can be found for most of the remaining terms. They, however, do not reflect the complete semantic meaning of the terms. In many cases, they do not capture the core semantic meaning either. Natural Language Processing (NLP) frequently uses standards for automated data retrieval and there is a need for mapping functional status descriptions to standard terminologies. Although standard terminologies may contain many of the important and useful concepts related to functional status, our study suggests that the use of NLP for functional status data extraction requires mapping ontologies to more descriptive terms and concepts such as the ones we identified.

While the percent coverage of functional status terms by standard clinical terminologies was similar comparing VINCI EMR to social media sources, there was a lack of overlap in the terms used across sources. This lack of exact overlap of terms between the VINCI EMR and social media online discussion forums reflects the differences in the context and purpose of the writing as well as their authors. Aside from paraphrasing patients’ reported functional status (e.g. “ambulate with assistance”) about their activities, EMRs tend to use medical language as well as templates for functional assessment instruments (e.g. Activities of Daily Living (ADLs), Functional Independence Measure (FIM)). Patients who describe their functional status on a social media website may tend to use their own language which can be more informal, erratic, expressive or ambiguous. It is worth noting that the mapping rates of the two sources are fairly similar. In other words, in this domain, the standard vocabularies represent neither the clinician nor the patient terminology well. In addition, no single vocabulary stood out as a particularly good source for functional status terms. The mapped terms were spread across a number of different terminologies.

We found several reasons for failures for the subset of non-matches and partial matches that we investigated (Table 4). While some concepts did not exist in the UMLS, many more failed to match completely due to use of a colloquial term, no match for a qualifier/modifier or too many descriptive terms. In the VINCI data some terms failed to match due to use of templates. For the social media data some terms were not completely matched due to use of metaphors.

Increasingly, researchers are examining standard terminologies in terms of their support of interoperability and coverage of concepts existing in the biomedical domain. For instance, Bodenreider et al. evaluated coverage by the UMLS for bioinformatics concepts from LocusLink and the Gene Ontology database. Others, including Frost et al. are developing sophisticated methods such as the Markov Chain Ontology Analysis to improve analytical applications for biomedical ontologies such as enrichment analysis, which quantifies the importance of ontology classes relative to a dataset.

However, to the best of our knowledge, we are aware of only one other study by Ruggieri et al. that examined the representation by standard terminologies of functional status concepts. The authors of this study found that the UMLS was superior to SNOMED-RT in matching functional status concepts from the Clinical Health Assessment Questionnaire (CLINHAQ) and the Modified Health Assessment Questionnaire (MHAQ). The CLINHAQ and MHAQ questionnaires are used to evaluate functional status of patients diagnosed with rheumatoid arthritis. Consistent with our study Ruggieri et al. reported that neither terminology had complete coverage of functional status terms and SNOMED-RT coverage was especially poor for concepts in the “activities” semantic class (terms found in the Physical Function and Role-Physical scales in our study).

There are a number of ongoing efforts to standardize functional status terminologies in industry. For example, standard representation of functional status has been actively studied in health information exchange and quality measures. Various groups including the Office of National Coordinator of Health Information Technology (ONC), Centers for Medicare and Medicaid Services (CMS) and Integrating the Healthcare Enterprise (IHE) are proposing functional status standards. The Consolidated-Clinical Document Architecture (C-CDA) specification defined the functional status section and specified vocabulary constraints for different types of functional status.
There has been increasing interest in automated extraction of functional status based on patient notes, including the use of the ICF standard terminology. One preliminary study, for example, used the ICF to extend an existing NLP system to encode functional status information noted in patient rehabilitation summaries. In our study ICF was among the UMLS terminologies that were examined. The coverage of functional status terms by ICF was low, 1.9% for the VINCI data and 4.1% for social media data. We note that terms in the ICF are generally noun phrases, however as Ruggieri et al. observed with functional-status language in questionnaires, we found that functional status information was mainly expressed with verbal phrases or sentences and included fewer medical terms. This may partly account for the low coverage.

One limitation of our study is that we focused on data from patients with atrial fibrillation or other cardiovascular diseases. Focusing on these patients may have limited the number of terms assigned to the Role-Emotional and Social Functioning SF-36 Health Survey scales. Future studies are needed to extend the research of functional status representation for other clinical domains. A second limitation is that we reviewed documents for a limited number of patients.

To address the gap in vocabulary coverage, we plan to update the Consumer Health Vocabulary (CHV), which is a UMLS source vocabulary developed by our research group, with terms identified in this study. Other possibilities include creating an ontology to represent the variety of functional status concepts and exploring the use of post-coordination. In anticipation of these efforts, one of the challenges we will face is to correctly map or link the terms, which are subjective by nature, to existing medical concepts. For example, it is not clear whether terms like “can’t get out of bed” or “feels like I’m going to die” refer to a physical or mental issue. At the same time they do indicate that the patients are not functioning well.

Conclusions

Individuals describe their physical status in a much more diverse way than current standard vocabularies capture. We collected a large number (2,763) of patient-reported functional status terms from both clinician and consumer sources of data, including terms from social media. With the increasing emphasis on patient-reported outcomes and the application of automated data retrieval methods, it is important to capture both patients’ own reports as well as those rephrased by clinicians. We further examined the coverage by current standard terminologies of functional status terms extracted from the VINCI EMR and social media data sources, sources that to our knowledge have not been previously examined. Overall, we found that standard terminologies do not sufficiently provide coverage of patient-reported functional status, and could be enhanced by studying these existing textual sources from clinicians and patients.

Acknowledgements

This work is funded by the US Department of Veterans Affairs, Office of Research and Development, Health Services Research and Development grants CHIR HIR 08-374, HIR 08-204, CRE 12-315 and the CREATE: A VHA NLP Software Ecosystem for Collaborative Development and Integration. Dr. Mohanty is supported by the VA Advanced Fellowship Program in Medical Informatics of the Office of Academic Affiliations, Department of Veterans Affairs. We would also like to acknowledge the staff, resources and facilities of the VA Salt Lake City IDEAS Center.
References


Reproducing a Prospective Clinical Study as a Computational Retrospective Study in MIMIC-II

Fabício S. P. Kury, MD, Vojtech Huser, MD, PhD, James J. Cimino, MD
National Library of Medicine, Bethesda, MD, United States

Abstract
In this paper we sought to reproduce, as a computational retrospective study in an EHR database (MIMIC-II), a recent large prospective clinical study: the 2013 publication, by the Japanese Association for Acute Medicine (JAAM), about disseminated intravascular coagulation, in the journal Critical Care (PMID: 23787004). We designed in SQL and Java a set of electronic phenotypes that reproduced the study’s data sampling, and used R to perform the same statistical inference procedures. All produced source code is available online at https://github.com/fablury/paamia2015. Our program identified 2,257 eligible patients in MIMIC-II, and the results remarkably agreed with the prospective study. A minority of the needed data elements was not found in MIMIC-II, and statistically significant inferences were possible in the majority of the cases.

Introduction and objectives
The presently widening scale of production and sharing of electronic health records (EHRs) [2] increases the interest on possible secondary uses for them. In this paper, we approached one publicly available dataset of electronic health records – the MIMIC-II database [3], version 2.6 – and applied one possible study design, namely, a computational retrospective study [4].

The central objective of our retrospective study was to demonstrate how MIMIC-II could be used to reproduce, and thereby validate, a contemporary prospective clinical study. The chosen prospective clinical study was by Gando et al., for the Japanese Association for Acute Medicine (JAAM) Sepsis Registry Study Group, published in 2013 in the journal Critical Care Medicine [1], hereafter referred to as the reference study. The objective of the reference study was to validate the prognostic value of the JAAM scoring system for diagnosis of disseminated intravascular coagulation (DIC). It was performed in Japan, from June 1, 2010 to May 31, 2011, and largely succeeded in demonstrating, among other things, that the higher the JAAM DIC score of the patient on day 1 of diagnosis of severe sepsis, the higher the mortality and morbidity. Therefore, our central objective was to reproduce the reference study in the MIMIC-II data, as an inexpensive computational, retrospective, observational study.

Scoring systems for disseminated intravascular coagulation
Disseminated intravascular coagulation (DIC) is a health condition characterized by the intravascular activation of the coagulation cascade [5] that is secondary to a range of disparate causes such as sepsis, pancreatitis, malignancy, heat stroke, and others. Prospective clinical studies have demonstrated that the development of DIC in patients with sepsis or severe trauma roughly doubles the risk of death, and is an independent predictor or mortality [7]. While its treatment hallmark remains to treat the provoking cause, there is interest in improving the identification of poor prognosis in DIC patients so that a more intensive or specialized therapy can be started earlier [8] or later [16] in the progression of the disease. The definition of DIC itself is not free from debate, and four scoring systems for the diagnosis of this condition have been published in the clinical literature by four institutions [5]. In this paper, we, intentionally replicating the reference study, applied and compared the one created and revised by the Japanese Association for Acute Medicine, and the one by the International Society of Thrombosis and Hemostasis (ISTH) (Table 1). Of note, our source code, available online [9], presently allows the computation of these two scoring systems as well as the other two (KSTH and JMHLW), which are outside the scope of the present study.

<table>
<thead>
<tr>
<th>SIRS* criteria</th>
<th>Japanese Association for Acute Medicine (JAAM)</th>
<th>International Society of Thrombosis and Hemostasis (ISTH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥3</td>
<td>+1</td>
<td>+2</td>
</tr>
<tr>
<td>0 to 2</td>
<td>0</td>
<td>+1</td>
</tr>
<tr>
<td>Platelet count</td>
<td>&lt;80 × 10⁹/l or &gt;50% decrease within 24 hours</td>
<td>≥50 &lt;100 × 10⁹/l</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≥100 × 10⁹/l</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elevated fibrin-related marker*</td>
</tr>
</tbody>
</table>
Reproducible research
To increase transparency and to allow re-execution of our analysis on other datasets, we voluntarily sought to adhere to the criteria for reproducible research as accepted by the journal Biostatistics [10]. Therefore, we offer the complete source code produced for this paper free online [9], while remarking that all data we used – MIMIC-II v. 2.6 – is publicly available for research, for free, via its own regulations. [3] We welcome critical review, correspondence and contributions to our work.

Methods
The reference study specified a list of data elements that were collected from patients in days 1 and 4 of diagnosis of severe sepsis, including the calculated JAAM and ISTH scores for DIC, and then used that data for statistical analysis. Our retrospective study had to therefore emulate these procedures in MIMIC-II as closely as possible. We used the Java and R languages because of our expertise with them and their free availability. The process is described below.

A. Data retrieval
We analyzed the reference study for its data sampling procedures and definitions, which Gando et al. dutifully report in the Data Sampling section of their paper, and manually listed all conceptual data elements it produced from each eligible patient (Table 2). Each item in that list would become the header of a column in a table called Clinical Data Table.

For knowing or calculating the value of those conceptual data elements, we knew, from their definitions, that we needed a much larger list of raw data elements. For example, for calculating the JAAM DIC score (a conceptual data element) it is necessary to know raw data elements from the patient such as platelet count, fibrinogen level, prothrombin time, and others. We manually produced a list of all those raw data elements needed, and then searched for the MIMIC-II-specific code(s) (ITEMIDs) for each. We did those searches inside MIMIC-II itself using SQL queries, and on the documentation of the database [11]. Table 3 brings examples of those raw data elements.

B. Electronic phenotyping
We used Java to reproduce the reference study’s inclusion criteria and data sampling procedures. Our program screened all hospital admissions of non-neonates in the database, and read the results of the laboratory tests and charted data in chronological order. For each case where the patient met the eligibility criteria, the program created one row in the Clinical Data Table, as defined in section A, and populated it with the required study data elements. In cases where MIMIC-II did not contain the required input data at the required point in time, which are common in retrospective studies [4], the Java program generated “Not Available” (NA) as output. The program eventually produced a CSV file that was imported into R for subsequent statistical analysis.

C. Statistical analysis
The reference study provided clear descriptions of the statistical analyses, which were reproduced using R. Groups were compared using Welch’s t-tests for data with unequal variances and Mann–Whitney U tests for data with equal variances. Effect of independent variables predicting mortality was calculated by stepwise logistic regression. For comparing patient survival, we simply counted how many patients were alive at the end of the follow-up period.

Results

A. Data retrieval

We identified 15 data elements produced by the Data Sampling methodology of the reference study, listed in Table 2. A few of them could potentially be derived from others, but were kept separate for simplicity. In order to produce all of them for each patient, we identified 49 raw data elements from MIMIC-II. Most of the raw data elements were MIMIC-II-specific ITEMIDs of charted data such as laboratory test results, but others were demographic or related to workflow such as patient age, gender, hospital admission date and date of death if it happened. Table 3 shows examples of raw data elements; the complete list can be found online together with this paper’s accompanying source code.

Table 2: All conceptual data elements required from each patient by the reference study

<table>
<thead>
<tr>
<th>Data element</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Number</td>
</tr>
<tr>
<td>Gender</td>
<td>Male/Female</td>
</tr>
<tr>
<td>Septic shock on day 1?</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Positive blood culture on day 1?</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Acute Physiology and Chronic Health Evaluation II score on day 1</td>
<td>Number</td>
</tr>
<tr>
<td>Sequential Organ Failure Assessment score on day 1</td>
<td>Number</td>
</tr>
<tr>
<td>Multiple Organ Dysfunction Syndrome (MODS) on day 1*</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Systemic Inflammatory Response Syndrome criteria on day 1</td>
<td>Number</td>
</tr>
<tr>
<td>JAAM DIC score on day 1</td>
<td>Number</td>
</tr>
<tr>
<td>JAAM DIC score on day 4</td>
<td>Number</td>
</tr>
<tr>
<td>ISTH DIC score on day 1</td>
<td>Number</td>
</tr>
<tr>
<td>Patient deceased within 28 days of diagnosis of severe sepsis?</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Patient deceased within one year of diagnosis of severe sepsis?</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Patient deceased during the hospital stay?</td>
<td>Yes/No</td>
</tr>
</tbody>
</table>

*Defined as SOFA score >= 12, as per the reference study.

Table 3: Examples of raw data elements required from MIMIC-II v. 2.6

<table>
<thead>
<tr>
<th>Data element</th>
<th>SQL Table</th>
<th>ITEMID(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>ADMISSIONS</td>
<td>-</td>
</tr>
<tr>
<td>Date of death if it happened</td>
<td>D_PATIENTS</td>
<td>-</td>
</tr>
<tr>
<td>Platelet count</td>
<td>LABEVENTS</td>
<td>50428</td>
</tr>
<tr>
<td>D-dimer plasma level</td>
<td>LABEVENTS</td>
<td>50370</td>
</tr>
<tr>
<td>Fibrin/fibrinogen degradation products plasma level</td>
<td>LABEVENTS</td>
<td>50376</td>
</tr>
<tr>
<td>PT (seconds) in blood</td>
<td>LABEVENTS</td>
<td>50439</td>
</tr>
<tr>
<td>PT (INR) in blood</td>
<td>LABEVENTS</td>
<td>50399</td>
</tr>
<tr>
<td>Temperature (Celsius) of blood</td>
<td>CHARTEVENTS</td>
<td>676, 677</td>
</tr>
<tr>
<td>Temperature (Fahrenheit) of blood</td>
<td>CHARTEVENTS</td>
<td>678, 679</td>
</tr>
<tr>
<td>Breath rate</td>
<td>CHARTEVENTS</td>
<td>3337</td>
</tr>
<tr>
<td>Heart rate</td>
<td>CHARTEVENTS</td>
<td>211</td>
</tr>
<tr>
<td>Overall SOFA score</td>
<td>CHARTEVENTS</td>
<td>20009</td>
</tr>
<tr>
<td>APACHE II Score</td>
<td></td>
<td>Not available</td>
</tr>
</tbody>
</table>

B. Electronic phenotyping

The eligibility criteria of our reference study consisted of the presence of diagnosis of severe sepsis [1]. Although not specified in the paper, we assumed as an additional, implied eligibility criterion that the patient be not a neonate. We based this assumption on the expectation that neonatal patients would have been mentioned in the reference study if they were included, since the usual setting in hospitals is to have separate ICUs for neonates and for non-
neonates. In MIMIC-II, a specialized variable — ICUSTAY_AGE_GROUP — informs you whether the patient is an “adult”, “neonate”, or “middle”. The eligibility criteria of our retrospective study consisted thus of the patient not being labeled “neonate”, and a diagnosis of severe sepsis be established.

Table 4 shows the definitions we used for identifying Systemic Inflammatory Response Syndrome (SIRS), sepsis, severe sepsis and organ failure from raw clinical data elements such as white blood cell count, partial pressure of carbon dioxide in arterial blood, urinary output in the last 24 hours, and others. They derive from the original definitions used in the reference study [13][14]. We translated our definitions into algorithms in Java that read each data element and compared it to its threshold(s) (e.g. Temperature < 36 C or >38.3C = one SIRS criteria).

**Table 4.** SIRS, sepsis and severe sepsis diagnostic criteria

<table>
<thead>
<tr>
<th>SIRS: 2 or more of the following:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Temperature &lt; 36 C or &gt; 38.3 C.</td>
</tr>
<tr>
<td>• Heart rate &gt; 90 beats per minute or &gt; 2 SD for age.</td>
</tr>
<tr>
<td>• Breath rate &gt; 20 breaths per minute.</td>
</tr>
<tr>
<td>• Abnormal white blood cell count:</td>
</tr>
<tr>
<td>• &gt; 12,000/µL,</td>
</tr>
<tr>
<td>• or &lt; 4,000/µL,</td>
</tr>
<tr>
<td>• or &gt; 10% immature neutrophil (band) forms.</td>
</tr>
</tbody>
</table>

**Sepsis**: SIRS + known or documented infection.

- Hospital admission must contain ICD-9 code for sepsis or septicemia.

**Severe sepsis**: Sepsis + organ dysfunction.

- Overall SOFA score ≥ 3.

Due to unavailability of ready means to know, from the data available in MIMIC-II, the moment when a patient first had a known or documented infection, which is a requirement for establishing a diagnosis of sepsis [13], in our electronic phenotype we adopted an alternate criterion for diagnosing severe sepsis. We accepted that, given the satisfaction of the remaining criteria of severe sepsis (Table 4), the requirement of known or documented presence of infection was satisfied by the presence of an ICD-9 code for either sepsis (ICD-9 = 995.91 or 995.92 [severe sepsis]) or septicemia (ICD-9 = 038.X) on the hospital admission. Our electronic phenotyping algorithm for severe sepsis thereby consisted as follows. First, SQL queries retrieved patients from MIMIC-II by their ICD-9 code for sepsis, severe sepsis or septicemia. For each of those patients individually, the Java program ingested and processed all pertinent EHR data. If the program found together two or more SIRS criteria, with at least one Overall SOFA score ≥ 3 in the same day, the patient was deemed eligible for our retrospective study, starting from the chart time of the data element with the latest chart time.

**Table 5.** Study eligibility screening results

<table>
<thead>
<tr>
<th></th>
<th>Reference study [1]</th>
<th>MIMIC-II Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total patients screened</td>
<td>14,417</td>
<td>27,579*</td>
</tr>
<tr>
<td>Total eligible patients</td>
<td>624</td>
<td>2,257</td>
</tr>
<tr>
<td>Gender (male/female)</td>
<td>391/233</td>
<td>1281/975**</td>
</tr>
<tr>
<td>Eligible patients with JAAM DIC diagnosis on day 1 (% of eligible patients)</td>
<td>292 (46.8)</td>
<td>395 (17.5)</td>
</tr>
<tr>
<td>28-day mortality in patients with/without JAAM DIC diagnosis on day 1 (%)</td>
<td>31.2/16.0</td>
<td>44/27.5</td>
</tr>
</tbody>
</table>

*: All non-neonates in MIMIC-II v. 2.6. Please note that this number differs from earlier versions of MIMIC-II.

**: For one patient that information was not available (NA).

For identifying groups of data elements together (e.g. the presence of two SIRS criteria together with a SOFA score above 2), we had to implement the concept of time windows, because each data element is charted at a potentially
different time. In a prospective study there is no necessity for such windows, because patient data (e.g. laboratory tests, vital sign measurements) is collected on-demand, according to the study protocol, virtually all at the same planned time. In our retrospective study, the essential question to be answered was: how far from each other, in time, can patient data elements be (e.g. a platelet count, a fibrinogen level measurement, and prothrombin time measurement) and still be considered effectively simultaneous?

The definitions of the DIC scoring systems also assume that all needed tests are performed on-demand, virtually at the same time, thus they also do not provide such specifications of time windows. Therefore, to adapt the prospective study’s design to a retrospective study design, we arbitrated two separate time windows.

- **Time window one:** for answering whether one data element, such as one record of overall SOFA score, was present at a point in time T, the algorithm returned the record closest to that time T, in the future or in the past, within 24 hours of distance of T, that is, within the interval T ± 24 hours.

- **Time window two:** for the calculation of scores of DIC or SIRS, which are aggregations of raw data elements, at a given point in time T, we considered only the EHR data charted until that time T and within the 48 hours immediately before T. That is, we considered only data charted within the interval from “T − 48 hours” until T.

Inside any time window, and for any data element, the record of that data element that was closest to time T was always the one used.

Table 6 exposes the characteristics of the patients after the screening for eligibility and calculation of DIC scores, reproducing the composition of Table 3 of the reference study. The difference is the column “Missing data”, which informs the percentage of patients that, on day 1, did not have that data element available.

**Table 6. Characteristics of JAAM DIC and non-DIC patients on the day of inclusion (day 1)**

<table>
<thead>
<tr>
<th></th>
<th>JAAM DIC (n = 395)</th>
<th>Non-DIC (n = 1565)</th>
<th>P value</th>
<th>Missing data (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>59.84 ± 16.84</td>
<td>66.95 ± 16.39</td>
<td>&lt;0.0001</td>
<td>0</td>
</tr>
<tr>
<td>Gender (male/female)</td>
<td>214/181</td>
<td>925/640</td>
<td>0.0790†</td>
<td>0.03</td>
</tr>
<tr>
<td>JAAM DIC score</td>
<td>4.74 ± 1.22</td>
<td>1.26 ± 0.52</td>
<td>&lt;0.0001</td>
<td>13.16††</td>
</tr>
<tr>
<td>Platelet count (× 10^9/l)</td>
<td>103.98 ± 101.56</td>
<td>224.83 ± 132.03</td>
<td>&lt;0.0001</td>
<td>0.1</td>
</tr>
<tr>
<td>Prothrombin time (seconds)</td>
<td>16.28 ± 4.2</td>
<td>15.67 ± 4.85</td>
<td>0.0125</td>
<td>1.48</td>
</tr>
<tr>
<td>Prothrombin time ratio</td>
<td>1.87 ± 1.24</td>
<td>1.7 ± 1.05</td>
<td>0.0122</td>
<td>2.45</td>
</tr>
<tr>
<td>Fibrinogen (g/l)</td>
<td>4.4 ± 2.44</td>
<td>4.38 ± 2.1</td>
<td><strong>0.9139</strong></td>
<td>61.94</td>
</tr>
<tr>
<td>FDP (mg/l)</td>
<td>91.81 ± 181.01</td>
<td>53.48 ± 152.85</td>
<td><strong>0.0624</strong></td>
<td>83.21</td>
</tr>
<tr>
<td>SIRS score</td>
<td>2.42 ± 0.67</td>
<td>2.37 ± 0.66</td>
<td><strong>0.1720</strong></td>
<td>0</td>
</tr>
<tr>
<td>SAPS I score</td>
<td>16.82 ± 5.5</td>
<td>16.31 ± 5.19</td>
<td><strong>0.1093</strong></td>
<td>9.03</td>
</tr>
<tr>
<td>SOFA score</td>
<td>10.72 ± 4.61</td>
<td>8.27 ± 3.89</td>
<td>&lt;0.0001</td>
<td>0</td>
</tr>
<tr>
<td>MODS (%)</td>
<td>41.5%</td>
<td>22.4%</td>
<td>&lt;0.0001</td>
<td>0</td>
</tr>
<tr>
<td>28-day mortality (%)</td>
<td>44.1%</td>
<td>27.5%</td>
<td>&lt;0.0001</td>
<td>0</td>
</tr>
<tr>
<td>1-year mortality (%)</td>
<td>59.7%</td>
<td>49.1%</td>
<td>0.00013</td>
<td>0</td>
</tr>
</tbody>
</table>

* The count of patients add up not to the number of total eligible patients in the study, but only to the number of patients whose DIC status was known on day 1, that is, 13.16% less than all eligible patients.
† From a t-test comparing the proportion of male/female between the two groups.
†† This cell differs from the others in the column because here the denominator is the number of eligible patients, rather than the number of eligible patients with known DIC status.
Figure 1, below, illustrates how the JAAM DIC diagnosis on days 1 and 4 correlate with mortality and Multiple Organ Dysfunction Syndrome (MODS). It is possible to identify a relative drop in the numbers in the case where the patients did not have JAAM DIC on day 1, then acquired it by day 4, which differs from this same graph in the reference study.

Figure 1. Mortality and Multiple Organ Dysfunction Syndrome (MODS) in JAAM DIC patients

The charting time of calculated scores – SIRS and DIC

The specific points in time of any score assessment or diagnosis (and suspension of thereof) naturally coincided with the chart times of the data in MIMIC-II. We considered the chart time to be the moment when, ideally, each new piece of data about the patient data became known in the original clinical setting that originated MIMIC-II, such as a new platelet count, or a new measurement of body temperature. All patient data was read and processed in chronological order according to the chart time.

C. Statistical analysis

Our results agreed with those of the reference study. As the JAAM DIC score at day 1 increased, so did the mortality at day 1, as well as the SAPS I and organ dysfunction scores, as can be seen in Table 7. One intriguing exception was the score of 7, which did not follow the trend.

Table 7. Disease severity, organ dysfunction and mortality for JAAM DIC score on inclusion day (day 1)*

<table>
<thead>
<tr>
<th>JAAM DIC score on day 1</th>
<th>1 (n = 1224)</th>
<th>2 (n = 276)</th>
<th>3 (n = 65)</th>
<th>4 (n = 254)</th>
<th>5 (n = 71)</th>
<th>6 (n = 8)</th>
<th>7 (n = 42)</th>
<th>8 (n = 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAPS I score</td>
<td>16.18 ± 5.12</td>
<td>16.77 ± 5.61</td>
<td>16.74 ± 4.47</td>
<td>16.32 ± 5.21</td>
<td>17.41 ± 5.51</td>
<td>18.14 ± 5.21</td>
<td><strong>17 ± 6.39</strong></td>
<td>20.21 ± 6.21</td>
</tr>
<tr>
<td>SIRS score</td>
<td>2.31 ± 0.63</td>
<td>2.59 ± 0.71</td>
<td>2.4 ± 0.75</td>
<td>2.33 ± 0.62</td>
<td>2.68 ± 0.73</td>
<td>2.62 ± 0.74</td>
<td><strong>2.26 ± 0.63</strong></td>
<td>2.9 ± 0.72</td>
</tr>
<tr>
<td>SOFA score</td>
<td>8 ± 3.74</td>
<td>9.13 ± 4.17</td>
<td>9.72 ± 4.7</td>
<td>10.31 ± 4.28</td>
<td>10.79 ± 4.55</td>
<td>14.12 ± 5.46</td>
<td><strong>11.12 ± 5.34</strong></td>
<td>13.55 ± 5.7</td>
</tr>
<tr>
<td>SOFA peak during</td>
<td>10.05 ± 4.38</td>
<td>4.74 ± 5.43</td>
<td>4.74 ± 5.08</td>
<td>12.49 ± 5.08</td>
<td>12.82 ± 5.08</td>
<td>13.17 ± 5.48</td>
<td><strong>13.88 ± 5.01</strong></td>
<td>16.95 ± 5.38</td>
</tr>
<tr>
<td>hospital stay</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8 shows that a diagnosis of DIC by the JAAM scoring system increased the odds of mortality by 18.2% to 43.2% (95% confidence interval), and that a decrease in the JAAM score within 4 days was associated with a 13.3% reduction in the odds of mortality (Table 8).

Table 8. Stepwise logistic regression analysis on day of inclusion (day 1) for prediction of 28-day mortality

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>P value</th>
<th>95% confidence interval</th>
<th>Missing data (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.020</td>
<td>0.00013</td>
<td>0</td>
</tr>
<tr>
<td>JAAM DIC score</td>
<td>1.300</td>
<td>&lt;0.0001</td>
<td>1.182 to 1.432</td>
</tr>
<tr>
<td>Delta JAAM DIC score (day 1 - day 4)</td>
<td>0.867</td>
<td>0.00516</td>
<td>0.784 to 0.958</td>
</tr>
<tr>
<td>Fibrinogen</td>
<td>0.998</td>
<td>0.00014</td>
<td>0.998 to 0.999</td>
</tr>
</tbody>
</table>

Discussion

Our endeavor in this paper builds upon a central concept that is as follows. Given structured data from the electronic health record, and cautious design of the data processing pipeline, a computer algorithm is able to calculate some popular scoring systems used in clinical practice. The criteria for the Systemic Inflammatory Response Syndrome is one example of such a scoring system, as well as essentially all four criteria for diagnosis of disseminated intravascular coagulation published in the literature [5]. Supported by this concept, we have accepted our electronic phenotype for severe sepsis without formal validation against a gold standard. Otherwise, one could interpret the high agreement of our results with those of the reference prospective study as evidence that the electronic phenotype, and the calculations of the used disease scores, had altogether acceptable performance.

Foremost, this paper shows the value of retrospective studies for clinical research, and that MIMIC-II can be used for them. Despite not a new resource [3], and the many studies with MIMIC-II [3], to our knowledge this is the first validation of a clinical study done in this valuable database. Our results (especially Tables 8 and 9) demonstrated resilience towards the limitations of a retrospective study, and remarkably followed the direction of the results of the reference study, which was done at a different country, and across 15 hospitals instead of one (which is the case of MIMIC-II).

Table 9. Comparison of final conclusions between the reference study and this study

<table>
<thead>
<tr>
<th>Sentence from the Results part of the Abstract of the reference study</th>
<th>Findings in our retrospective study</th>
<th>Agreement between studies?</th>
</tr>
</thead>
<tbody>
<tr>
<td>The prevalence of JAAM DIC was 46.8% (292/624)</td>
<td>Incidence of JAAM DIC on day 1 was 17.5% (395/2,257).</td>
<td>No.</td>
</tr>
<tr>
<td>and 21% of the DIC patients were scored according to the reduction rate of platelets.</td>
<td>Not assessed.</td>
<td></td>
</tr>
</tbody>
</table>
The JAAM DIC patients were more seriously ill and exhibited more severe systemic inflammation, a higher prevalence of multiple organ dysfunction syndrome (MODS) and worse outcomes than the non-DIC patients. Disease severity, systemic inflammation, MODS and the mortality rate worsened in accordance with an increased JAAM DIC score on day 1. Dynamic changes in the JAAM DIC score from days 1 to 4 also affected prognoses. The JAAM DIC score on day 1 (odds ratio = 1.282, P < 0.001) and the Delta JAAM DIC score (odds ratio = 0.770, P < 0.001) were independent predictors of 28-day death. Dynamic changes in the JAAM DIC score from days 1 to 4 also affected prognoses. The JAAM DIC scoring system included all patients who met the International Society on Thrombosis and Haemostasis (ISTH) overt DIC criteria on day 1. The International Society on Thrombosis and Haemostasis scoring system missed a large number of nonsurvivors recognized by the JAAM scoring system.

Limitations

**Missing data:** A problem inherent to a retrospective study, the extent of missing data in our study ranged from zero to 83.21% (Fibrin/Fibrinogen Deg. Prod. [FDP] level) or 100% (“known source of infection”). We did not perform statistical artifacts on the data to fill in missing values. As can be seen in Table 6, it was of utmost importance to this study the fact that the clinical data required by the reference study is commonly collected as part of routine care. The appraisal of data availability can easily be the foremost concern to a researcher considering a retrospective study. For the “known source of infection”, which we deemed completely unavailable and translated into the presence of an ICD-9 code for sepsis or septicemia in the patient’s hospitalization data, this workaround can be expected to be imparting selection bias according to imperfections in the ICD-9 coding. For the FDP level, as well as other charted patient information, our program understood as Not Available (NA) any needed data element that was not available within the requested time window. This approach brought the design
caveat that for aggregations of raw data elements such as SIRS and DIC scores, it was not clear how to differentiate a known zero from a NA. This is because any pertinent data element can cause each score to be nonzero. However, in a retrospective study, it can easily be too restrictive to require that all possible data elements must be available before you can accept a known zero. This is because, unlike in a prospective study, you cannot collect patient data on-demand to satisfy your data needs. Our approach to this issue was a simplifying one – we chose to ignore any score below 1, and treat them as Not Available. Our understanding was that, because the diagnostic threshold is much higher than 1 (Table 1), this would be of minimal to no impact to our study. The absence of score zero can be readily noticed in Table 7.

**Diagnosis of severe sepsis/DIC near death:** From manual inspection of the data, we noticed that in a considerable (approx. 10%) of the cases a diagnosis of severe sepsis and/or DIC was only ever made in the last hours before the moment of patient death – and not rarely a bit after the time of death. We rationalized that, near the moment of passing, a clear laboratorial diagnosis of severe sepsis or disseminated intravascular coagulation is clouded by a greater disruption of body homeostasis, and the computation is clouded by possible imperfections in data collection such as tests collected or charted after the moment of death. Therefore, we ruled to completely ignore any patient data charted within 12 hours of the time of patient death. Patients could be electronically phenotyped only before that point in time.

**Issues with data representation:** Three data elements from MIMIC-II – namely: D-Dimer level, Fibrin/Fibrinogen Degradation Products level, and Fibrinogen level – were found to be available as notations rather than only numbers, for example: `<10`, `10-40`, `160-320`, `>10000`. We decided that our Java program would thereby perform conversion of those values.

- The sign `<` was interpreted as `95%`, thus `<10` became 9.5;
- ranges were averaged, thus `160-320` became 240;
- The sign `>` was interpreted as `105%`, thus `>10000` became 10500.

Furthermore in this issue, some scoring systems requested data in a level of granularity that was not available in MIMIC-II. That was the case of Fibrin/Fibrinogen Degradation Products (FDP) level for the JAAM DIC scoring system. The JAAM DIC scoring system defines three ranges – 0-10, 10-25 and >25 – with different values, however, that data element is available in mimic as the following ranges: 0-10, 10-40, 40-80, and so on. Our referred averaging method turned `10-40` into 25 (3 points in the JAAM DIC) and `0-10` into 5 (0 points), effectively impeding any patient to receive just 1 point for that criteria inside the JAAM DIC.

**Conclusion**

We reproduced a contemporary prospective clinical study as a computational retrospective study in an EHR database – MIMIC-II – which found considerably more eligible patients than the prospective study, and was executed at a fraction of the cost and time. Despite the limitations of the retrospective study design, the results demonstrated resilience and agreed with those of the prospective study in a large majority of the findings.

Our results speak in favor of the usefulness of computational retrospective studies [4] as eventual proxies to costlier clinical studies, and suggest an altogether good performance, in MIMIC-II, of our developed electronic phenotypes for SIRS, severe sepsis and two scoring systems for disseminated intravascular coagulation available in the clinical literature.

We strived to maximize the dissemination and reuse of the work by making all source code we produced available online free [9], in addition to the fact that MIMIC-II is also available free to the research community.

From the clinical perspective, we offer our results as one contribution to the body of evidence on the validity and use of scoring systems for diagnosis of disseminated intravascular coagulation. In particular, we highlight the value of our study as a positive third-party validation of the work by Gando et al.
Acknowledgements
This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and Lister Hill National Center for Biomedical Communications (LHNCBC). This research was also supported in part by an appointment to the NLM Research Participation Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy (DoE) and the NLM.

Conflicts of interest
The authors declare no conflicts of interest in this study.

References
Computer-Supported Feedback Message Tailoring for Healthcare Providers in Malawi: Proof-of-Concept

Zach Landis-Lewis¹, Gerald P Douglas¹, Harry Hochheiser¹, Matthew Kam², Oliver Gadabu¹, Mwatha Bwanali³, Rebecca S Jacobson¹
¹Center for Health Informatics for the Underserved, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA; ²Google, New York, USA; ³Baobab Health Trust, Lilongwe, Malawi

Abstract

Although performance feedback has the potential to improve the quality and safety of care, healthcare organizations generally lack knowledge about how this guidance is best provided. In low-resource settings, tools for theory-informed feedback tailoring may enhance limited clinical supervision resources. Our objectives were to establish proof-of-concept for computer-supported feedback message tailoring in Malawi, Africa. We conducted this research in five stages: clinical performance measurement, modeling the influence of feedback on antiretroviral therapy (ART) performance, creating a rule-based message tailoring process, generating tailored messages for recipients, and finally analysis of performance and message tailoring data. We retrospectively generated tailored messages for 7,448 monthly performance reports from 11 ART clinics. We found that tailored feedback could be routinely generated for four guideline-based performance indicators, with 35% of reports having messages prioritized to optimize the effect of feedback. This research establishes proof-of-concept for a novel approach to improving the use of clinical performance feedback in low-resource settings and suggests possible directions for prospective evaluations comparing alternative designs of feedback messages.

Introduction

Globally there are significant gaps between best practices drawn from medical evidence and decisions made by healthcare professionals⁶. Closing these gaps is increasingly difficult because of accelerating rates of the production of biomedical knowledge and the increasing complexity of healthcare systems⁷. The urgency of overcoming these challenges has motivated the development of a Learning Health System (LHS), a cyber-social infrastructure that enables both learning from data and the creation of feedback loops to improve the delivery of patient-centered care⁸.

Audit and feedback (AF), defined as the provision of clinical performance summaries to healthcare providers, teams, and organizations, is widely used to support learning and behavior change for healthcare quality improvement⁹. A recent Cochrane review including 140 clinical trials shows that AF can significantly improve compliance with desired practice, but that it is unclear how and when it is effective⁰. Seeking to improve the utility of AF evidence, researchers have called for a shift from overall effectiveness studies towards comparative effectiveness studies, evaluating how and when AF intervention components will workⁱ. To better understand AF components, researchers have argued for the explicit use of psychological theory that explains how AF operates to change behavior and support clinical learning⁵. Frameworks that enable the use of theory in behavior change interventions include the Theoretical Domains Framework (TDF) and the Capability, Opportunity, Motivation and Behavior (COM-B) system¹¹. A mapping between the TDF and COM-B enables researchers to model determinants of behavior within COM-B, and to relate these determinants to a broader set of causal mechanisms within theoretical constructs in the TDF that may hold implications for the effective delivery of performance feedback in organizations.

We have proposed that a theory-informed feedback tailoring tool used by clinical supervisors could improve AF for individual healthcare provider recipients¹². Computer-supported feedback tailoring requires clinical supervisors to use their familiarity with the recipient and situation to select an optimal feedback message from a menu of potentially relevant messages. This approach may be especially useful in low-resource settings where factors such as staff turnover, disruptions to care, and minimal dataset collection create a need to handle high uncertainty and missing information in the analysis of clinical data¹³. Understanding the design space for individually-tailored feedback messages is a key step toward designing and implementing such a system. Our objectives for this research were to analyze clinical performance data to understand the requirements and potential impact of computer-supported feedback message tailoring in a low-resource setting.
Methods

To establish proof-of-concept for using computer-supported feedback message tailoring in low-resource settings, we used a five-stage process (Figure 1). First, we measured clinical performance using four guideline-based indicators. Second, we developed a preliminary model of the effect of feedback on performance for a range of barriers to performance improvement in antiretroviral therapy (ART) clinics. Third, we created a rule-based message tailoring process that used the model of the effect of feedback in our specific context. Fourth, we used the message tailoring process to retrospectively generate menus of tailored messages about each individual healthcare provider within the performance dataset. Finally, we analyzed the performance dataset and the resulting tailored message data to understand the potential impact for computer-supported feedback message tailoring in this context.

Figure 1. Study design to establish proof-of-concept for a computer-supported feedback message tailoring system

Setting and data collection

We evaluated our approach in Malawi, where healthcare providers use a national electronic medical record system (EMR) in the provision of ART in public hospitals\(^\text{4}\). Malawi is a landlocked country in Sub-Saharan Africa with a population of close to 17 million people. The country has a largely agricultural economy, a highly rural population, and high rates of poverty, with approximately 74% of the population earning less than $1.25 per day. Like most low-income countries, Malawi has a significant shortage of healthcare providers. With a ratio of approximately one physician for every 50,000 inhabitants, Malawi and neighboring Tanzania have the lowest doctor-to-patient ratio in the world\(^\text{4}\). For this reason, care in ART clinics is primarily provided by clinical officers (non-physician clinicians with 3 to 4 years of post-secondary medical training) and nurses. We collected de-identified EMR data from ART clinics in public hospitals in Malawi. This research was approved by the University of Pittsburgh Institutional Review Board (IRB), protocol #PRO12100159 and the Malawi National Health Sciences Research Committee (NHSRC), protocol #1019.

Performance measurement

Performance indicators are commonly used to identify potential problems that may represent data quality problems or valid exceptions to recommended clinical practice\(^\text{5}\). In prior research we developed a method for identifying guideline-based performance measures that could be routinely evaluated within EMR data for ART in Malawi\(^\text{6}\). Using this approach, we identified four performance measures based on statements from Malawi’s national guideline for the clinical management of HIV, 2011 edition\(^\text{7}\) which have also been used for quality improvement purposes in multiple Sub-Saharan African countries\(^\text{8}\) (Table 1). We measured individual-level rather than clinic-level performance because each task is commonly performed independently and because performance may be influenced by individual differences in knowledge, skills and motivation.

The performance measures in Table 1 use ratios that accommodate both individual and team-based care. The denominator reflects the total number of opportunities a provider had to provide recommended care to each patient. For example, if a provider used the EMR to document an ART visit with a patient who was eligible to receive cotrimoxazole preventative therapy (CPT), this patient was counted towards the total number of patients in the provider's denominator for the month of that visit. The numerator reflects the documented care received by the patients who were counted in the individual provider’s denominator, regardless of who provided the care to the patient. For example, in a scenario where a provider does not prescribe CPT to an eligible patient at the time of an ART visit, but the patient receives a prescription for CPT on another day during the review period from any other provider, the patient would still be counted in the provider’s numerator. One exception is for patients whose care happens to be provided adequately but over a time frame spanning review periods. We used the Ruby programming language, the MySQL database system, and R statistical analysis software to measure and graph performance for each of the four performance indicators. To validate the results we reviewed the scripts, queries, and graphs of the
performance data with EMR developers. We conducted the review by discussing the approach for performance measurement, the structure of each SQL query, and the definitions of clinical concepts used by each query.

Table 1. Guideline-based performance indicators for ART in Malawi17,18,19

<table>
<thead>
<tr>
<th>Performance indicator</th>
<th>Malawi ART guideline recommendation</th>
<th>Numerator</th>
<th>Denominator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitoring of nutritional status: Pediatric patient height</td>
<td>“Record length / height to the nearest cm at every visit (children)” (2011 edition, page 18)</td>
<td>Number of children with height recorded at least once during the review period</td>
<td>Number of children with at least one clinical visit during the review period</td>
</tr>
<tr>
<td>Monitoring of nutritional status: Weight</td>
<td>“Record weight in kg to the nearest 100g at every visit” (2011 edition, page 18)</td>
<td>Number of patients with weight recorded at least once during the review period</td>
<td>Number of patients with at least one clinical visit during the review period</td>
</tr>
<tr>
<td>Cotrimoxazole Preventative Therapy (CPT) prescribing</td>
<td>“Provide CPT to all patients in HCC and ART follow-up” (2011 edition, page 32)</td>
<td>Number of patients who were prescribed CPT</td>
<td>Number of patients with at least one clinical visit during the review period without CPT contraindications</td>
</tr>
<tr>
<td>WHO clinical staging</td>
<td>“WHO clinical staging is mandatory for all HIV patients” (2011 edition, page 12)</td>
<td>Number of patients with a WHO clinical stage at the time of ART initiation</td>
<td>Number of patients who were initiated on ART during the review period</td>
</tr>
</tbody>
</table>

Modeling the effect of feedback on clinical behavior

We created a preliminary model of the influence of feedback on ART performance, adapted from an earlier model of the effect of feedback on clinical behavior20 (Table 2). The use of this model requires several key assumptions to be met. First, we assume that performance is measured routinely at individual and clinic levels so that a recipient's performance can be interpreted in relation to group performance. Second, we assume that a supervising clinician is charged with giving feedback and has a) some awareness of the events that have occurred in the clinic during the performance measurement interval and b) some awareness of or willingness to make estimations about the individual recipients' determinants of behavior such as knowledge, skills, and motivation18. We demonstrate how the tailoring approaches in the rightmost column of Table 2 could be applied in a prototype feedback planning menu that a clinical supervisor to could use to select tailored messages for a healthcare provider (Table 3).

Message tailoring process

Using the performance features from Table 2, we developed a rule-based message tailoring process (Table 4) to identify and make inferences about observable features of clinical performance data. The purpose of the tailoring process is to infer when individualized feedback might be useful, in various formats, and to prioritize a set of tailored feedback messages within a menu (Table 3) for a clinical supervisor to use.

1. Identify performance features: The first step of the message tailoring process is to classify each known feature as present or absent for an individual's performance. Performance features are the individual and situational characteristics associated with an individual provider, and his or her behavior that is targeted by an AF intervention. Our objective was to demonstrate the feasibility of using a range of performance features that could support inferences about barriers to behavior change and their associated theoretical constructs. We selected a preliminary set of 11 performance features such as “consistently low performance” and “consistently high performance” based on our understanding of the clinical setting gained from a qualitative study of performance feedback in ART clinics in Malawi18. We anticipate that this feature set is a small sample of the meaningful features that could be used.

2. Infer construct salience: Once performance features are identified, these can be used to make inferences about an individual's potential barriers to behavior change and associated theoretical constructs. For example, when individuals demonstrate consistently high performance for a behavior, we may infer that barriers to behavior change are not likely to be present. Conversely, consistently low performance may indicate that barriers associated with capability constructs such as a lack of knowledge or skills are in effect, especially when accompanied by concurrent high performance of peers.
Table 2. A preliminary model of the influence of feedback on antiretroviral therapy (ART) performance

<table>
<thead>
<tr>
<th>COM-B Capability</th>
<th>TDF domain</th>
<th>TDF construct</th>
<th>Barrier to ART performance improvement</th>
<th>Hypothetical causal mechanism for individual feedback</th>
<th>Potential influence of feedback</th>
<th>Performance features</th>
<th>Tailoring approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>Knowledge</td>
<td>Awareness of guideline</td>
<td>Feedback can change awareness to impart new knowledge that leads to behavior change</td>
<td>High</td>
<td>Consistently low individual performance, no prior feedback provided</td>
<td>Current score: Prioritize individual feedback, include guideline recommendations</td>
<td></td>
</tr>
<tr>
<td>Procedural</td>
<td>Knowledge</td>
<td>Awareness of performance</td>
<td></td>
<td></td>
<td></td>
<td>Peer comparison: Prioritize individual feedback, include peer comparison</td>
<td></td>
</tr>
<tr>
<td></td>
<td>knowledge</td>
<td>Knowledge of how to use the EMR</td>
<td></td>
<td></td>
<td>Consistently low individual performance, no prior feedback provided</td>
<td>Current score: Prioritize individual feedback, recommend or provide EMR training</td>
<td></td>
</tr>
</tbody>
</table>

| Opportunity       | Social influences | Social pressure | Material resources | Social influences and resources based on availability of material resources | None (no direct influence on social norms) | Low | Low group performance | Withhold feedback: Withhold or deprioritize individual feedback |
| Social pressure   | Social norms     | Peer pressure and social norms | | | | |

| Motivation        | Beliefs about capabilities | Self-efficacy | Beliefs about specific capabilities | Feedback can influence perceptions of ability that lead to behavior change | Conditions on situation | Low individual performance, improvement trend | Self-comparison: Prioritize individual feedback using a truncated scale graph when improvement trend is present |

3. **Infer message component relevance:** Some components of potential messages will be more relevant to an individual and situation that others. For example, showing a graph of performance data that emphasizes a trend by using a truncated scale will only be relevant when the improvement trend is above some threshold of meaningful improvement. We propose that performance features can also be used to determine which elements of feedback messages are most relevant for a recipient. For example, performance features could be used to estimate the relevance of the following feedback message components from the tailored messages in Table 3:

- Scale truncation: The use of a truncated vertical axis to emphasize change in performance
- Self comparison: Comparing an individual’s past performance with current performance
- Peer comparison: Comparing an individual’s current performance with current peer performance.

4. **Prioritize messages:** Finally, we prioritized messages appearing in the menu to place messages that were most likely to address barriers to behavior change at the top of the menu. To prioritize the messages, we evaluated both the message component relevance scores and estimated construct salience scores for the provider’s current month of performance. We created preliminary rules based on hypothetical causal mechanisms from Table 2 to prioritize each of the tailoring approaches in Tables 2 and 3 (Current score, peer comparison, withhold feedback, and self-comparison). We created two additional prioritization categories: no prioritization for the cases where no messages could be prioritized, and prioritized combination, in the event that two or more messages were equally of highest priority.
Table 3. Example prototype feedback planning menu for ART weight recording performance

<table>
<thead>
<tr>
<th>Beliefs about determinants of recipient's weight recording behavior</th>
<th>Tailoring approach</th>
<th>Proposed design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recipient may be unaware that the ART guidelines recommend recording a patient's weight at each clinic visit.</td>
<td><strong>Current score:</strong> Prioritize individual feedback, include guideline recommendations</td>
<td>In the last month, 22% of patients who you provided care for in the ART clinic had their weight recorded. Malawi’s national ART guideline recommends that all patients have their weight recorded at each ART visit (2011 edition, page 18).</td>
</tr>
</tbody>
</table>
| Recipient may be unaware of how he or she is performing relative to peers. | **Peer comparison:** Prioritize individual feedback, include peer comparison | Your performance: 22%
Top peers: 85%
In the last month, your performance was more than 60% below your top-performing peers in the ART clinic. |
| Recipient may not know how to use the EMR to record a patient's weight. | **Current score:** Prioritize individual feedback, recommend or provide EMR training | In the last month, 22% of patients who you provided care for in the ART clinic had their weight recorded. To record a patient's weight using the EMR, select “Record vitals” from the ART visit menu... |
| Recipient does not have access to a functioning scale to record patient weight. | **Withhold feedback:** Withhold or deprioritize individual feedback | [None] |
| A common practice in the clinic is to deprioritize the recording of patient's weight relative to other tasks, especially during times of high workload. | **Withhold feedback:** Withhold or deprioritize individual feedback | [None] |
| Recipient is aware of his or her past performance. Recipient may believe he or she is not capable of improving performance. Increased effort is likely to lead to improvement. | **Self-comparison:** Prioritize individual feedback using recent performance history and a truncated scale graph when improvement trend is present | In the last 2 months, your weight recording performance has improved by 7% |

*Analysis of performance data and prioritized messages*

We conducted two simple analyses to understand the potential impact of message tailoring in ART clinics in Malawi: identification of performance gaps, and analysis of variability in message prioritization. We first calculated the average number of performance gaps occurring each month to estimate how frequently peer comparison feedback could be provided. We defined a performance gap as a 10% or greater difference in performance between an individual healthcare provider and the average performance of two higher-performing peers working in the same clinic in the same month. We excluded providers who had seen 10 or fewer patients in a month.
Table 4. Feedback message tailoring process

<table>
<thead>
<tr>
<th>1. Identify performance features</th>
<th>2. Infer construct salience</th>
<th>3. Infer message component relevance</th>
<th>4. Prioritize messages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method and inputs</strong></td>
<td>Feature classification using performance data</td>
<td>Rule-based inference using performance features</td>
<td>Rule-based inference using performance features</td>
</tr>
<tr>
<td><strong>Output data type</strong></td>
<td>Binary (Present/absent)</td>
<td>Integer score</td>
<td>Integer score</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td><strong>Feature:</strong> consistently_low_performance</td>
<td><strong>Construct:</strong> self-efficacy Rule: If consistently_low_performance is present, increase the salience score for self-efficacy as a barrier to improvement</td>
<td><strong>Message component:</strong> peer_comparation Rule: If a 10% performance gap exists between the recipient and top-performing peers, increase the relevance score for messages containing peer_comparation</td>
</tr>
<tr>
<td><strong>Rule:</strong> If performance has remained below 50% during the performance interval then consistently_low_performance is present</td>
<td></td>
<td></td>
<td><strong>Rule:</strong> If self_efficacy is salient as a barrier, decrease the priority score for messages using peer_comparation</td>
</tr>
</tbody>
</table>

Our second analysis concerned the variability of prioritized messages within a tailoring menu. The purpose of this analysis was to understand how often a standardized message format could be used for all providers as compared to the routine tailoring of messages. We anticipate that increased variability of message priorities indicates a greater potential impact of message tailoring because it reflects individual and situational differences among healthcare providers that standardized feedback formats are less likely to accommodate. If messages of the same format have the same priority for more than 95% of providers, it would suggest that message tailoring is not necessary. However, if the priority of messages is more evenly stratified across message type groups, and if the size of these groups changes over time, it would suggest a greater potential impact for message tailoring. We created a prioritized list of message types for each performance measure across all individuals and months. To assess the variability of message formats, we identified the top-priority message format only and did not consider the rank of lower-priority message formats. We calculated the percentage of individual performances that had each of the four types of message as the highest priority, plus the two additional prioritization categories.

**Results**

*Data collection*

Data collection for this study occurred in November, 2013. We collected de-identified EMR data from 11 ART clinics in Malawi that were using the National ART EMR (Software version BART 1), and which was recorded between October, 2011 and September, 2013.

*Performance measurement*

We retrospectively measured clinical performance from 372 healthcare providers in 11 hospital-based ART clinics over a two-year period. We measured individual performance for all four measures at a monthly frequency by individual provider for a total of 7,448 individual monthly performance reports having a denominator of five or more patients. Across all clinics, an average of seven automated monthly performance reports could be generated per month for each performance indicator.

Summary statistics for each of the four guideline-based performance measures for all healthcare providers grouped by ART clinic are shown in Table 5. The average monthly performance for weight recording and WHO clinical staging were consistently high (> 94%) for all but one clinic. The average monthly performance for pediatric height recording ranged from 2.3% to 98.4%, while the average monthly performance for CPT prescribing ranged from 48.9% to 87.6%. For CPT prescribing performance, there was a generalized decrease in performance in 2012 across clinics, with providers at most clinics having a wide range of performance during the period. The generalized decrease in performance is associated with a national shortage of CPT drugs that occurred in 2012.
Performance gaps of 10% or greater

Performance gaps that could be used for peer-comparison feedback occurred slightly more frequently than 1 gap per month on average across all clinics and measures. The mean monthly total of performance gaps for each indicator across sites between October, 2011 and September, 2013 is shown in Figure 2. The average number of performance gaps that could be used to give peer comparison feedback to a single provider for all 11 sites ranged from 0.32 to 2.45 gaps per month.

![Graph showing performance gaps](image)

Figure 2. Average monthly performance gap totals between October, 2011 and September, 2013

Message prioritization variability

Based on the preliminary rule set, the tailoring process resulted in 35% (2,624 / 7,448) of individual monthly reports being prioritized to optimize the effect of feedback on performance. We calculated the percentage of all messages that had each message type as the highest priority on an individual’s monthly report (Figure 3). No reports had peer comparison messages as the highest priority. Across all performance indicators, increased stratification of tailored message types appears to be associated with lower performance. For example, the indicators having higher performance, which are weight recording and WHO clinical staging, had a higher average percentage of messages that were not prioritized, at 75% for weight and 83% for WHO clinical staging. In contrast, pediatric height recording and CPT prescribing, which have lower overall performance, had increased stratification of highest priority percentages across message types.
Discussion

We found that computer-supported feedback message tailoring could be routinely used in ART clinics in Malawi. The results of this study answer several important questions about using EMR data to generate tailored performance feedback messages in a low-resource setting. Most significantly, we identified an opportunity to use existing EMR data to routinely monitor individual clinical performance and provide tailored feedback across a range of guideline-based performance indicators in a low-resource setting. This approach could be expected to yield individualized monthly reports for ART providers working at each site, with approximately 35% of reports being tailored to optimize the effect of feedback on performance. Although performance appears to allow limited room for improvement in some ART clinics, we found regular opportunities to provide individualized feedback to address performance gaps and potential performance or data quality problems. These findings are significant because the existing National EMR infrastructure in Malawi would allow these reports to be generated in every ART clinic using the EMR, totaling 66 clinics at the end of June, 2015. Moreover, such a system may enable feedback to be generated more rapidly than the current quarterly reporting schedule of the National ART monitoring and evaluation program.

We sought to understand if feedback message tailoring could potentially impact clinical performance by exploring differences in features of performance data. We found differences in performance features that appear to hold meaningful implications for the design of feedback messages. On average, based on a preliminary set of hypothetical causal mechanisms offered by behavioral and cognitive theories, more than 50% of feedback messages for pediatric height recording could be tailored for individual or situational differences in performance. Similarly, close 50% of feedback messages could be tailored for differences in performance with regard to CPT prescribing in this setting. Where performance is higher, there appear to be fewer opportunities to tailor feedback messages. However, even the indicators having higher performance allowed for routine tailoring for approximately 25% of messages for weight recording on average, and for an average of 16.3% of messages for WHO clinical staging.

These findings are significant because they represent the first evaluation, to our knowledge, that uses a model of clinical behavior and psychological theory for the purpose of feedback message tailoring. To our knowledge, this approach represents a novel contribution that holds implications for related research in biomedical informatics, implementation science, and global health. In the field of biomedical informatics, this work introduces a novel class of knowledge-based system to support evidence-based care and quality improvement. In implementation science, this work is the first demonstration of theory-informed tailoring of feedback messages for individual healthcare providers based on standardized measures of clinical performance. In the domain of global health, this work represents the first supervision tool of its kind for a setting where supervision resources are limited. The significance of these findings increases with the increasing availability of eHealth data that can be used to generate performance feedback.

This research has several limitations. A key limitation of this analysis is that the application of theory within tailoring rules was not rigorously validated. In a prior model formation study we applied theory using knowledge gained from a review of the literature. Furthermore, the model we developed for the effect of feedback on ART performance was a general model that did not accommodate potential differences in barriers to behavior change.
between the four clinical behaviors. Developing behavior-specific models, which is an important next step, may yield different message tailoring approaches.

A potential limitation of this approach is the ability of a clinical supervisor to accurately perceive barriers to behavior change that inform the selection of feedback messages. We anticipate that supervisors who use a feedback message tailoring system could overcome this limitation by observing the effects of their selected feedback messages on clinical performance. These observations could enable a supervisor to learn how to improve the use of feedback for individual recipients.

Our assessment did not account for data quality problems, however we anticipate that performance feedback could also be used to target the improvement of clinical data quality where such problems exist. The message tailoring process may also provide a systematic approach for integrating routine data quality assessment into AF, to inform clinical supervisors when clinical data is not “fit for use” as performance feedback.

The classification thresholds that we used for this analysis were chosen based on our understanding of the clinical context rather than empirical research. For example, we classified low group performance as an average performance below 50%, but it is likely that the actual thresholds for low group performance may vary across ART clinics, and across performance indicators. In the case of WHO clinical staging, a threshold for low group performance might be set much higher for most clinics because there are no valid exceptions for this guideline recommendation. In the case of CPT prescribing there are exceptions for patients with CPT contraindications, therefore a lower threshold would be expected. To address this limitation, we chose classification thresholds that err on the side of a lower bound, meaning that thresholds that we validate are likely to lead to greater variability of tailored messages.

Finally, a limitation of this proof-of-concept approach is that the manual development and maintenance of a message tailoring process for each clinical context is not likely to be sustainable. We aim to address this and other limitations in future research by developing a publicly-maintained feedback message tailoring knowledge-base that builds on established frameworks for applying psychological theory to behavior change interventions\textsuperscript{20,21}, and which uses an argumentation model to represent dynamic and contextualized evidence\textsuperscript{22}.

Conclusion

Computer-supported feedback message tailoring is a promising approach for improving the use of AF within LHSs. In this research we explored the design space for individually tailored feedback messages, to establish proof-of-concept for computer-supported feedback tailoring in Malawi. We found that individually-tailored performance feedback can be generated using routinely collected EMR data in ART clinics in Malawi. This finding suggests that clinical supervisors could use feedback tailoring tools to improve the effect of feedback on clinical performance in low-resource settings. Future research should study the use of a feedback tailoring system and its impact on clinical performance.

References

Contrasting Association Results between Existing PheWAS Phenotype Definition Methods and Five Validated Electronic Phenotypes

Joseph B Leader, BA1, Sarah A Pendergrass, PhD MS2, Anurag Verma, MS2,4, David J Carey, PhD3, Dustin N Hartzel, BS1, Marylyn D Ritchie, PhD MS2,4, H. Lester Kirchner PhD1

1Center for Health Research, 2Biomedical and Translational Informatics Program, 3Weis Center for Research, Geisinger Health System, Danville, PA, USA; 4The Center for Systems Genomics, The Pennsylvania State University, University Park, PA USA

Abstract

Phenome-Wide Association Studies (PheWAS) comprehensively investigate the association between genetic variation and a wide array of outcome traits. Electronic health record (EHR) based PheWAS uses various abstractions of International Classification of Diseases, Ninth Revision (ICD-9) codes to identify case/control status for diagnoses that are used as the phenotypic variables. However, there have not been comparisons within a PheWAS between results from high quality derived phenotypes and high-throughput but potentially inaccurate use of ICD-9 codes for case/control definition. For this study we first developed a group of high quality algorithms for five phenotypes. Next we evaluated the association of these and 4,636,178 genetic variants with minor allele frequency > 0.01 and compared the results from high-throughput associations at the 3 digit, 5 digit, and PheWAS codes for defining case/control status. We found that certain diseases contained similar patient populations across phenotyping methods but had differences in PheWAS.

Introduction

Phenome-Wide Association Studies (PheWAS) are used to investigate the association between a collection of genetic variants and a wide and diverse range of phenotypes, diagnoses, traits, and/or outcomes. PheWAS highlight novel connections between multiple phenotypes and can elucidate more of the phenotype-genotype landscape as well as generate new hypotheses for further exploration. The complex results of PheWAS also have the potential for uncovering new mechanistic insights. Electronic health record (EHR) data coupled with genetic data have been used repeatedly for PheWAS. For example, through the Electronic Medical Records and Genomics (eMERGE) network, EHR based International Classification of Diseases, Ninth Revision (ICD-9) codes have been used to establish case/control status for identifying significant associations between medical record diagnoses and genetic data, including using datasets from the Geisinger Health System (GHS) MyCode™ biorepository. Each PheWAS has shown the ability to replicate previously reported associations, clearly showing that the high-throughput PheWAS approach is valid as well as identifying hypothesis-generating novel associations.

EHR based PheWAS have used various methods of grouping and counting ICD-9 codes to define case/control status for analyses. For example, there are three digit ICD-9 codes that specify disease categories (e.g. code 405 for “secondary hypertension”) that can be further subdivided using multiple four or five digit sub ICD-9 codes (e.g. 405.1 for “benign secondary hypertension”, 405.11 “benign renovascular hypertension”), that are more specific. The most common methods of defining case/control status using EHR data include grouping ICD-9 codes by three digits (e.g. 250.02 rolls up to 250), five digits (e.g. 250.02 stays 250.02), and “PheWAS Codes”. The choice of which patients to consider a case or control based on their number of instances of a given code and who should be excluded (not classified as case or control) for a code has varied depending on the study. For example, one approach defines case status by requiring three or more instances of a given code per individual (rule of three), considering those with no instances of the code a control, and one or two instances of a code are excluded from further analysis. PheWAS Codes go a step further by collapsing and grouping ICD-9 codes that are highly related and expanding the exclusion of individuals based on ICD-9 codes with revisions based on code frequency and human review.

Case and control status based on the presence of an ICD-9 code has been one of the most accessible forms of EHR-based high-throughput phenotypes allowing researchers to look for known and novel associations across a very
broad range of phenotypes. This can be attributed to the fact that the phenotype definitions can be easily applied across the spectrum of ICD-9 codes and results are more easily compared to other EHR based PheWAS. These methods have limitations, however; the positive and negative predictive values (PPV/NPV) vary greatly from phenotype to phenotype. For example, a single diagnosis of myocardial infarction is highly predictive whereas a single diagnosis of stroke is not highly predictive.

While existing studies have demonstrated that PheWAS is a viable method to identifying genome/phenome associations which replicate well known associations, the measurement of EHR based PheWAS success have been focused on known associations and not a comparison of the phenotype definitions themselves. Thus, for this study, using the Geisinger Health System MyCode Biorepository, we identified ICD-9 code based case/control status using multiple methods. We then compared the results of the existing phenotype algorithms case/control status to high-quality algorithmically defined EHR phenotypes ("gold standard") case/control status for type 2 diabetes (T2DM), acute coronary syndrome (ACS), non-obstructive coronary artery disease (CAD), obstructive CAD, and obesity. Finally, we performed a genome-wide association study for each phenotype and contrasted highly significant results of these GWAS with association results from performing PheWAS using different ICD-9 code based case/control definitions.

Methods

Study Participants

The MyCode dataset had a total of 3,022 individuals available with both phenotypic and genotypic data for this study. Because the majority of subjects were of European ancestry (EA), only EA subjects were selected for these analyses.

Gold Standard Algorithmically Defined Phenotypes

Five phenotype algorithms developed within Geisinger Health System with a high PPV/NPV (≥ 95%) were selected to compare to the phenotype definitions from various approaches for identifying case/control status based only on ICD-9 codes. These phenotype algorithms are considered to be “gold-standard” because of their high PPV. The algorithms for these gold standard phenotypes incorporate EHR data (labs, meds, procedures, etc.) and temporality in addition to ICD-9 codes used in traditional PheWAS phenotyping. Table 1 illustrates the different types of data used in the gold standard phenotypes and shows the gold-standard phenotype overlaps with ICD-9 codes for different ICD-9 code levels/abstractions. The asterisks (*) in the table refer to a wildcard operator which would indicate any digit (0-9) included in the range.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Data Entities Used in Gold Standard Algorithm</th>
<th>PheWAS Codes</th>
<th>3 Digit</th>
<th>5 Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type II Diabetes Mellitus</td>
<td>ICD-9 Codes, Medications, Laboratory Results</td>
<td>250.2</td>
<td>250*</td>
<td>250.0, 250.2</td>
</tr>
<tr>
<td>Acute Coronary Syndrome</td>
<td>ICD-9 Codes, Laboratory Results, Cardiac Catheterization Results, Radiology Results, ECG Results</td>
<td>411, 411.2, 411.4, 411.9</td>
<td>410*, 411*</td>
<td>410.* 411.1</td>
</tr>
<tr>
<td>Non-Obstructive Coronary Artery Disease</td>
<td>ICD-9 Codes, Laboratory Results, Procedures, Interventional Radiology Results</td>
<td>411, 411.3</td>
<td>413*</td>
<td>413.*</td>
</tr>
<tr>
<td>Obesity</td>
<td>ICD-9 Codes, Laboratory Results, Procedures, Interventional Radiology Results</td>
<td>278.<em>, 513.</em></td>
<td>278.*</td>
<td>278.*</td>
</tr>
<tr>
<td>Obstructive Coronary Artery Disease</td>
<td>ICD-9 Codes, Laboratory Results, Procedures, Interventional Radiology Results</td>
<td>411, 411.3</td>
<td>413*</td>
<td>413.*</td>
</tr>
</tbody>
</table>

At Geisinger Health System, data for creating phenotype algorithms are extracted from Geisinger’s Clinical Decision Intelligence System (CDIS), an enterprise data warehouse, additionally supplemented data from the Department of Cardiology. CDIS is updated every 24 hours with feeds from multiple source systems, including the EHR, tumor registry, financial decision support, claims, patient satisfaction and high-use third-party reference.
datasets. The source data are transformed through a complex “Extract-Transform-Load” (ETL) process. Laboratory results, medications, procedures, past medical history, and encounter setting are often key components of validated phenotypes and help to improve PPV/NPV.

Phenotype algorithm development is a very iterative process that incorporates input from researchers, clinicians, clinical support staff, and Information Technology (IT) teams to understand how and why data was captured at the point of care in order to inform algorithm development. Programmer/Analysts in the Phenomic-Analytics and Clinical Data Core (P-ACDC) use existing methods of phenotyping and expand the process to allow for increased iteration and validation. First, analysts receive input from the members of the research team to determine if all data needed for the phenotype exists in the data warehouse and if not, the analyst will work with IT staff to incorporate that data into CDIS. Once complete data is obtained to define a phenotype, the algorithm is developed. Using the algorithm, one hundred positive (cases) and 100 negative patients (controls) for that phenotype are identified, and an analyst reviews the clinical records to validate the algorithm and calculate the PPV/NPV. If the predictive values are below desired levels (ie, < 85%), the analyst will identify areas of improvement, modify the algorithm, and repeat the validation process until desired results are achieved.

All attributes of the validated phenotypes are saved as value sets (terminology) and business rules (setting, timing, etc.). The collection of value sets and business rules is based on existing processes and repositories so that phenotypes can be consumed and shared across organizations. The National Library of Medicine (NLM) Value Set Authority Center (VSAC) (https://vsac.nlm.nih.gov/), the Center for Disease Control and Prevention (CDC) Public Health Information Network Vocabulary Access Distribution System (PHIN VADS) (http://www.cdc.gov/phin/tools/PHINvads/index.html), Phenotype Knowledge Base (PheKB) (https://phekb.org/), and PhenotypePortal (http://phenotypeportal.org/) are well known national standards for consuming and authoring value sets and business rules.

Matching ICD-9 Based Phenotypes to Gold Standard Phenotypes

ICD-9 codes from the MyCode participants EHR that also had genotypic data were used for this study. For our 3 digit and 5 digit PheWAS analyses, case status was defined as having ≥ 3 visits per individual, zero instances of an ICD-9 code to be considered a control. Individuals between 0 and 3 instances of a code (i.e., 1-2 visits) were omitted from the analyses for that code. At least ten case subjects per ICD-9 code to retain that ICD-9 code in association testing was required.

Genotypic Data and Quality Control

GHS MyCode subjects were genotyped using the Illumina HumanOmniExpress-12 v1.0 array a total of 729,078 SNPs. Imputation was used to improve genomic coverage of the datasets. The imputation was performed using the IMPUTE2 algorithm after phasing with SHAPEIT211 using the 1,000 Genomes cosmopolitan reference panel, resulting in a total of 38,054,243 SNPs in 3,111 samples for MyCode12. Genotype Quality Control (QC) procedures were performed prior to association testing using the R programming statistical package13 and PLINK software14. The first step was to filter out the SNPs with poor imputation quality; SNPs with imputation quality scores > 0.9 were used for further analyses. Data were filtered further for 99% genotype and sample call rates and minor allele frequency (MAF) threshold of 1%. Also, related samples were removed using Identity by Descent (IBD) kinship coefficient estimates. After QC, the genotypic data consisted of 4,636,178 SNPs and 3,022 samples with both phenotypic and genotypic data from MyCode.

Association Testing

Logistic regression with an additive encoding for the SNPs was used to evaluate the association between SNPs and the gold standard phenotypes. For each association, models were adjusted for decade of birth and sex. Association results were filtered for p-values < 1x10^-5, to focus further analyses on highly significant results associated with the gold standard variables. Results were annotated using Biofilter15, to add information about any genes that the SNPs were in or near, as well as to annotate the SNPs with any known results from the NHGRI GWAS catalog16. The NHGRI GWAS catalog contains results from published GWAS in the literature reaching genome-wide significance.
Next, a PheWAS was performed using all SNPs identified through the previous step with \( p < 1 \times 10^{-4} \). As described above, logistic regression with an additive encoding of the SNPs was used while controlling for decade of birth and sex. This was performed at the 3 digit and 5 digit ICD-9 code diagnosis level, and with PheWAS codes. Results for the association testing were plotted using GGPLOT2\(^1\) and PheWAS-View\(^1\) software.

**Results**

**Comparing Case and Control Status: Gold Standard Phenotypes and ICD-9 Methods**

After patients were classified as either a case or control using the various methods a comparison was conducted to demonstrate the overlap existing between the defined case/controls sets according to each method (Table 2). Not surprisingly, it was found that the methods agree in some areas and not in others. For example, T2DM was very closely matched across all of the methods (Figure 1) and obesity showed a large overlap (Figure 2) using diagnosis codes but incorporating vital measurements in the gold standard picked up additional patients that were otherwise missed. The cardiovascular phenotypes did not show agreement between the 3-digit and 5-digit code methods with the gold standard because of the inherent lack of information in the ICD-9 codes themselves. Using results from the cardiac catheterization laboratory and elevated biomarkers helps to appropriately classify the disease state. Overall, evaluating the absolute value of the difference in the number of cases according to the gold standard definition with that of PheWAS Codes, 3 digit codes, and 5 digit codes, it is observed that 3 digit and 5 digit codes have the closest case counts to the gold standard in all phenotypes except for obesity. To better understand the various case classifications, Venn diagrams were used to show the differences in actual patient overlap, not just absolute count.

### Table 2. Counts of patients identified by phenotype method.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Gold Standard</th>
<th>PheWAS Code</th>
<th>3 digit</th>
<th>5 digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2DM</td>
<td>1012</td>
<td>1029</td>
<td>1033</td>
<td>1024</td>
</tr>
<tr>
<td>ACS</td>
<td>158</td>
<td>870</td>
<td>129</td>
<td>74</td>
</tr>
<tr>
<td>Non-Obstructive CAD</td>
<td>208</td>
<td>870</td>
<td>134</td>
<td>134</td>
</tr>
<tr>
<td>Obesity</td>
<td>1741</td>
<td>1334</td>
<td>1321</td>
<td>1283</td>
</tr>
<tr>
<td>Obstructive CAD</td>
<td>315</td>
<td>870</td>
<td>134</td>
<td>134</td>
</tr>
</tbody>
</table>

**Figure 1.** Venn diagram demonstrating the overlap of cases by phenotype method for T2DM

**Figure 2.** Venn diagram demonstrating the overlap of cases by phenotype method for Obesity
Figure 3. Venn diagram demonstrating the overlap of cases by phenotype method for ACS

Comparing Association Results between Gold Standard Phenotypes and ICD-9 Code Based PheWAS

A total of 2,585 SNPs passed the p-value cutoff for the gold-standard phenotypes. Eight of these SNPs were present in the NHGRI GWAS Catalog. Those SNPs significantly associated with gold-standard phenotype at p-value > 1x10^{-4} were evaluated by performing a PheWAS with the ICD-9 codes at 3 digit, 5 digit diagnosis level, and PheWAS code level. Table 3 shows the average p-value of the gold standard associations passing our p-value cutoff, and what the maximum p-value was. Table 3 also shows for each gold standard phenotype the number of associations with p < 1x10^{-4} matching the same SNP-phenotype gold association. Note, more than one ICD-9 code corresponds to some of the gold standard phenotypes (Table 1), thus matches were only counted once when they occurred. The results of the PheWAS using different ways to define case control status recapitulate what was observed in the case/control definition comparisons of Table 2. For phenotypes where the ICD-9 codes offer more specific or appropriate terminology about a diagnosis, such as with T2DM, there is a better match in associations identified through PheWAS. For phenotypes poorly represented in ICD-9 code terminology compared to high-quality algorithmic development using multiple inclusion and exclusion criteria, the number of associations repeated in PheWAS compared to gold standard results is reduced, particularly in ACS.

Also, of note, while PheWAS Codes had higher case numbers than the algorithms for CAD, this did not result in more overlap in associations with the gold standard associations when compared to the 5-digit and 3-digit approaches. Ranges of p-values for associations where there was the same SNP from the gold standard phenotype and a matching ICD-9 code from Table 1 were evaluated and shown in Table 4. Depending on the gold standard phenotype, one case/control approach (PheWAS codes, 5 digit, and 3 digit) performs better than another method, however the differences are mostly minimal across the definitions of case/control status. While the case numbers can be lower for ICD-9 codes compared to gold standard algorithms, which can affect power to detect associations, there will also be more individuals incorrectly defined as cases and controls when using ICD-9 based case/control definitions alone. As a result, this will impact the ability to detect significant associations, as well as the degree of the significance of the association. Supplementary materials have containing all the individual association results used for Table 3 and 4 are located online: https://ritchielab.psu.edu/publications-supplementary-data/amia-phewas.

Table 3. Range of gold standard p-value results and number of associations for the same SNP and phenotype matched across three methods of defining ICD-9 case/control status.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Gold p-value avg</th>
<th>Gold p-value max</th>
<th>Number Gold Phenotype SNPs</th>
<th>Matched PheWAS Code Associations</th>
<th>Matched 5 Digit Associations</th>
<th>Matched 3 Digit Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2DM</td>
<td>5.33E-05</td>
<td>1.00E-04</td>
<td>563</td>
<td>559</td>
<td>558</td>
<td>562</td>
</tr>
<tr>
<td>ACS</td>
<td>5.17E-05</td>
<td>9.99E-05</td>
<td>399</td>
<td>13</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>Non-Obstructive CAD</td>
<td>3.80E-05</td>
<td>9.97E-05</td>
<td>738</td>
<td>114</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td>Obesity</td>
<td>5.03E-05</td>
<td>9.99E-05</td>
<td>310</td>
<td>109</td>
<td>149</td>
<td>90</td>
</tr>
<tr>
<td>Obstructive CAD</td>
<td>5.14E-05</td>
<td>1.00E-04</td>
<td>721</td>
<td>159</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>
Table 4. The range of p-values (minimum, average, maximum) for associations matching the same SNP and gold-standard defined phenotype across three methods of defining case/control status.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Phecode p-value min</th>
<th>Phecode p-value avg</th>
<th>Phecode p-value max</th>
<th>5 Digit p-value min</th>
<th>5 Digit p-value avg</th>
<th>5 Digit p-value max</th>
<th>3 Digit p-value min</th>
<th>3 Digit p-value avg</th>
<th>3 Digit p-value max</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2DM</td>
<td>1.01E-04</td>
<td>7.54E-04</td>
<td>9.81E-05</td>
<td>1.01E-04</td>
<td>7.54E-04</td>
<td>9.87E-05</td>
<td>1.00E-04</td>
<td>6.84E-04</td>
<td>9.80E-05</td>
</tr>
<tr>
<td>ACS</td>
<td>2.54E-04</td>
<td>5.88E-03</td>
<td>9.25E-03</td>
<td>1.16E-03</td>
<td>3.54E-03</td>
<td>9.10E-03</td>
<td>5.10E-04</td>
<td>3.28E-03</td>
<td>8.38E-03</td>
</tr>
<tr>
<td>Non-Obstructive CAD</td>
<td>1.59E-04</td>
<td>2.06E-03</td>
<td>9.93E-03</td>
<td>1.86E-03</td>
<td>4.76E-03</td>
<td>7.88E-03</td>
<td>1.87E-03</td>
<td>4.80E-03</td>
<td>7.94E-03</td>
</tr>
<tr>
<td>Obesity</td>
<td>1.05E-04</td>
<td>3.72E-03</td>
<td>4.56E-05</td>
<td>1.21E-04</td>
<td>3.25E-03</td>
<td>9.77E-06</td>
<td>1.07E-04</td>
<td>3.87E-03</td>
<td>1.83E-05</td>
</tr>
<tr>
<td>Obstructive CAD</td>
<td>1.59E-04</td>
<td>3.21E-03</td>
<td>3.70E-05</td>
<td>1.38E-03</td>
<td>5.58E-03</td>
<td>9.38E-03</td>
<td>1.40E-03</td>
<td>5.67E-03</td>
<td>9.51E-03</td>
</tr>
</tbody>
</table>

Figure 4. Manhattan-plots of all results. These are the results for the 5-digit ICD-9 code defined case/control diagnoses. Supplementary figures show the results for the 3-digit and PheWAS Code defined case/control status and are available online. The points in gold show the significance of the association result for the ICD-9 based diagnoses most similar to the gold-standard defined phenotype, for example the ICD-9 code 250.00 is the ICD-9 code for T2DM.

Variability in the identification and significance of associations was observed across the various methods of defining case/control status definitions when compared to the results using gold standard phenotypes. Some of this variability may be due to shifting case identification; different ICD-9 approaches resulted in identifying different individuals. Interestingly, the study did not observe the 3 digit or PheWAS code approach as having the most significant results of the three approaches from SNP to SNP, even though there could have been increased power through increased case numbers. Three examples are shown (Figure 5-7) of single SNPs. and the results across three different methods for the phenotypes of obesity, non-obstructive and obstructive CAD, and T2DM. Larger versions of figures are available in the online supplementary materials. For all three of these SNPs the ICD-9 diagnoses most similar to these gold-standard phenotypes have a similar direction of effect of the association. Figures 5-7 illustrate the significance of association through ordering by most significant at the top of the figure with decreasing significance plotted clockwise. The length of the each line corresponds to the −log(p-value) of each result.
**Figure 5.** Sun Plot of association results for SNP rs7127254, coded allele T, present across the three methods of identifying case/control status used for these analyses. This SNP was associated with the gold standard phenotype of obesity in our study with p-value $6.19 \times 10^{-7}$, and the array of metabolic syndrome comorbidities also associated with SNP notable. The most significant association for this SNP was with the 5 digit ICD-9 of T2DM.

![Sun Plot for SNP rs7127254](image)

**Figure 6.** Sun Plot of results for SNP rs2277251, coded allele T. This SNP was associated with the gold standard phenotype of T2DM with p-value $7.08 \times 10^{-7}$. SNP-phenotype associations ($p < 0.01$) that were also present for this SNP across the 3 methods of identifying case/control status. The spectrum of other comorbidities related to T2DM also associated with this SNP is notable.

![Sun Plot for SNP rs2277251](image)

**Figure 7.** Sun Plot of results for SNP rs10009355, coded allele T. This SNP was associated with the gold standard phenotype of obstructive CAD with p-value $8.24 \times 10^{-7}$. SNP-phenotype associations ($p < 0.01$) that were also present for this SNP across the 3 methods of identifying case/control status. Additional comorbidities related to CAD are associated with this SNP.

![Sun Plot for SNP rs10009355](image)
Discussion

Studies have shown that PheWAS can be a viable method for identifying known and/or novel associations in phenotypes and genotypes for hypothesis generation. The current methods used to define phenotypes allow for simple application of business rules to define a wide range of diseases. Defining gold standard methods for phenotypes is a laborious task that cannot be reasonably applied to all diseases. Comparing these methods of defining phenotypes shows that there are certain diseases where using code based case methods agree with more complex phenotype algorithms ("gold standard") while other diseases cannot use codes alone and require additional data from the EHR. For example, good agreement across the various methods was observed for T2DM, whereas ACS exhibited poor agreement due to the lack of specificity PheWAS codes.

For the gold-standard phenotypes, some associations were identified from previously published replicated results (annotated in the supplementary results of associations). The sample size was relatively small and thus the analysis had low power to identify new associations or replicate known associations. Therefore, for these exploratory analyses, a less stringent p-value cutoff of $1 \times 10^{-4}$ was used and p-values for multiple testing were not corrected for. Future directions include exploring these analyses within a larger genetic dataset linked to EHR based phenotypic data as well as repeating these analyses with a "rule of one" applied, where the presence of an ICD-9 code will indicate an individual is a control, instead of the "rule of three" applied here. Requiring less instances of an ICD-9 code for case/status will increase the number of controls. For some ICD-9 codes this may increase the number of false-controls; however it may increase the power of associations for ICD-9 codes that are rare and unlikely to have been assigned to a patient incorrectly.

Additional research needs to be done to compare the positive and negative predictive value of the ICD-9 code based method phenotypes and compare the patient populations that do not agree to potentially identify additional methods that are generalizable and can be added broadly to phenotype definitions. In addition, comparing more gold standard phenotypes to code based approaches would help to identify disease areas that have a high degree of agreement between the simple and complex approaches. Collaborating with existing efforts such as eMERGE, VSAC, and PHIN VADS would allow comparison to be completed at a much larger scale and across multiple institutions. Machine learning approaches are also a next step in developing phenotypes that perhaps can be semi-automated.

References


Building Structured Personal Health Records from Photographs of Printed Medical Records

Xiang Li, PhD¹, Gang Hu, MS¹, Xiaofei Teng, PhD¹, Guotong Xie, MS¹
¹IBM Research, Beijing, China

Abstract
Personal health records (PHRs) provide patient-centric healthcare by making health records accessible to patients. In China, it is very difficult for individuals to access electronic health records. Instead, individuals can easily obtain the printed copies of their own medical records, such as prescriptions and lab test reports, from hospitals. In this paper, we propose a practical approach to extract structured data from printed medical records photographed by mobile phones. An optical character recognition (OCR) pipeline is performed to recognize text in a document photo, which addresses the problems of low image quality and content complexity by image pre-processing and multiple OCR engine synthesis. A series of annotation algorithms that support flexible layouts are then used to identify the document type, entities of interest, and entity correlations, from which a structured PHR document is built. The proposed approach was applied to real world medical records to demonstrate the effectiveness and applicability.

Introduction
Personal health record (PHR) is an electronic application through which individuals can access, manage and share their health information¹, which can help individuals take a more active role in their own health, and provide decision support to assist patients in managing chronic diseases². Data collection is a base functionality of a PHR system, and it is time-consuming and error-prone for individuals to manually enter their PHR data¹. Hence, a more practical manner for collecting PHR data is to give individuals access to their own electronic health records (EHRs), which has been or is being implemented in some developed countries. For example, in the United States, many individuals can access and download their own EHR data using Blue Button⁴. And in England, all patients will have online access to the EHR data that their general practitioners hold by 2015⁵.

In China, even though PHR has been given much attention by health authorities and researchers⁶, at the present stage it is still difficult for Chinese individuals to access their EHR data, especially in underdeveloped areas. So far, China has not released a privacy standard like HIPAA⁷ (which mandates that patients can access their health records), and many Chinese individuals are regarded as lack of security and privacy concerns. Therefore, China health institutions tend to be conservative in giving patients access to their EHR data. Moreover, there are dozens of EHR vendors in China, and the EHR systems from different vendors widely vary due to the lack of implementation guidelines. So it is also difficult to develop unified interfaces for individuals to access their data in EHRs. Instead, individuals can easily obtain from hospitals the printed copies of their medical records, such as prescriptions, lab test reports, etc. This situation is not unique to China, but is widespread in many regions, especially in the developing countries.

Smartphones are very popular nowadays. China now has more than 500 million smartphone users, most of whom are accustomed to use phone cameras to take pictures of paper-based documents, including printed medical records, for the purpose of storage. Figure 1 shows some printed Chinese medical records, which were photographed using mobile phones. Because the data in photographed medical records are not structured, they cannot be directly used to build PHRs. A potential solution is to identify structured data in them using optical character recognition (OCR) techniques.

However, it is a challenging problem to build structured PHR from the photographs of printed medical records. First of all, the images are shot by ordinary mobile phones rather than high-quality scanners. Compared with the scanned images, they tend to have uneven light and shade, lower clarity, more noise, as well as paper skewing. Secondly, different types of medical records are involved to create PHRs. Even for a specific type, as shown in Figure 1, the layouts of printed medical records are not standardized in China, which may obviously vary in different time and different hospitals. Furthermore, various categories of entities, such as diagnoses, medications, test items names and values, are supposed to be extracted from medical records, where different types of characters can be observed, including Chinese characters, English characters, digits and punctuation. Existing OCR systems in clinical settings⁸,⁹,¹⁰ focus on the identification of structured clinical or privacy information from scanned documents, and do not consider photographed images. Moreover, these systems were designed for standardized documents, where the entities of interest are written in English and Arabic numerals. None of these systems can support unstandardized documents that may have very flexible layouts or have complex character sets as Chinese medical records do.
In this paper, we address these issues by proposing an approach to extract structured data from the photographs of printed medical records. Before OCR, several image processing algorithms, including denoising, binarization and deskewing, are performed to reduce the influence of low image quality of photographs shot by mobile phones. Multiple OCR engines are then applied to recognize text, and the results of the engines are synthesized to achieve a higher performance in recognizing characters of a complex character set. Furthermore, a series of flexible annotation algorithms that support diverse document layouts are designed to identify document types, entities of interest and entity correlations, using machine learning, pattern matching, and slot-based techniques. Based on these algorithms, we developed a system pipeline to automatically build structured PHR documents from printed medical records. The proposed approach has been applied to build PHRs from real world prescriptions and lab test reports, with high precision and sensitivity.

**Methods**

The goal of the proposed approach is to accurately extract entities of interest and entity correlations from photographs of printed medical records as many as possible. Therefore, both precision and recall (sensitivity) of the approach should be taken into account when designing the system and algorithms. The system pipeline and sample data flow are shown in Figure 2. A user first shoots a printed medical record using an ordinary mobile phone camera, and sends it to the system. Then the system performs image pre-processing on the photo. The processed image is sent to multiple OCR engines to build documents with recognized text, which are then resegmented and synthesized in the post-processing step. After that, the document type, entities of interest and entity correlations of the document are automatically annotated. The credible entities are adopted to create a structured PHR document, which is finally sent back to the user for further editing. In this study, two types of documents, prescriptions and lab test reports, were involved. And four categories of entities were focused on, including diagnosis names, medication names, test item names and test item values, which are critical to collect for building PHRs.

**Corpus**

The corpus of this study consists of 100 printed medical record documents (including 54 prescriptions and 46 lab test reports, see Figure 1) from 19 different hospitals in China, which were voluntarily provided by individuals. These documents were respectively photographed by several volunteer users using their own mobile phones, including Apple iPhone 4S, iPhone5C, iPhone 5S, LG Nexus 5, Samsung SCH-I779, Xiaomi HM Note, etc. All records were mainly written in Chinese, with many numerical values and a small amount of English words. The corpus includes 1476 entities to recognize, including 84 diagnosis names, 111 medication names, 650 test item names and 631 test item values. For evaluation purpose, these entities were manually annotated by three human readers.
Figure 2. The system pipeline and sample data flow
Image Pre-processing

Document photographs that are shot by mobile phone cameras tend to have lower quality than scanned images due to uneven light and shade, photo noise and skewing. For obtaining reliable OCR results, image pre-processing of the photographs needs to be first performed to enhance image quality.

To reduce the negative impacts of light and shade on OCR, image binarization, which converts an image into a black-and-white image, was performed. Since a photo shot by phone usually has different lighting conditions in different areas, the binarization algorithms that use a global value as threshold cannot generate a reasonable binary image. Thus, we applied the adaptive thresholding algorithm that calculates the local threshold for a small region of the image. For a pixel, its local threshold is the weighted sum of neighborhood grayscale values of the pixel, where the weights are computed in a Gaussian window. The pixel whose grayscale value is greater than the local threshold is assigned white, otherwise it is assigned black. Figure 3(c) shows a binarized image derived from the photo in Figure 3(a).

As shown in Figure 3(c), the binarized result of a photo shot by phone can have much salt-and-pepper noise, which may lead to additional OCR errors in some cases. Notice that most of the salt-and-pepper noise in the binarized image is derived from the Gaussian noise in the original photos. So we used the non-local means denoising algorithm, which replaces a pixel with the average color of the most resembling pixels in a search window, to remove the Gaussian noise in an original image before it is binarized (see Figure 3(b)). And as shown in Figure 3(d), a majority of the salt-and-pepper noise in the binarized image were removed accordingly. In this study, we utilized the implementation of the binarization algorithm and the denoising algorithm provided in the OpenCV 2.4.9 library.

Since it is difficult for users to exactly align their phone screens with paper-based documents during photographing, documents in photos often become skewed, which also reduces OCR accuracy. Therefore, we performed a deskewing algorithm that rotates an image to make text run as horizontally across the document as possible (Figure 3(e)). In this study, we utilized the deskewing algorithm in the image process toolkit provided by IBM Datacap Taskmaster Capture 8.1. We also tried some image cleanup algorithms such as line removal and dot shading removal offered by Datacap. However, in our experiments, these cleanup algorithms did not significantly improve the OCR performance because they also removed useful strokes in text. So these algorithms were not integrated in our current system pipeline.

Optical Character Recognition

For building PHRs, multiple types of characters were supposed to be recognized in medical records, such as Chinese characters, English characters, digits and punctuation. In general, a single OCR engine cannot achieve a satisfactory accuracy in recognizing characters of this complex character set. For example, the accuracy of a Chinese OCR engine in recognizing English characters and digits is usually relatively low, while an English OCR engine normally identifies Chinese characters as garbled text. To address this problem, we applied multiple OCR engines and synthesized the results of the engines in the following post-processing phase. In this study, the OCR was performed using IBM Datacap Taskmaster Capture 8.1, which embeds the Nuance OmniPage OCR engines and supports OCR in multiple languages. We applied the embedded Chinese OCR engine and the English OCR engine, and Figure 4(b) and 4(c) show the results of these two engines respectively for the processed image shown in Figure 4(a). The OCR errors of the examples are highlighted, including the character recognition errors (red circle) and the word segmentation errors (red blank). Note that here we show the results of the OCR engines in plain text for simplicity. Actually, the output of each OCR engine was formatted as a XML document with a line-word-character structure, where each line, word or character has its minimum bounding box B and a confidence value v ∈ [0, 1].
As mentioned above, the results of the multiple OCR engines were synthesized in the post-processing phase. Before doing that, the word segmentation errors from the OCR engines should be corrected, because both the synthesis and annotation approaches work on word-level, and incorrectly segmented words can cause extraction errors.

OCR engines normally segment characters into words based on inter-character intervals, and rarely consider the syntax information such as punctuation. For example, non-numeric characters separated by a full point “.” should be segmented, while two integers separated by a point “.” (e.g., “5.50”) usually represent a decimal number that should not be split. As these criteria are not captured by the OCR engines, we performed two rules sequentially to resegment the words more accurately: 1) Punctuation segmentation, which splits words using a set of pre-defined punctuation, including English punctuation such as “.” and “,” as well as Chinese punctuation such as “。” and “，”; 2) Decimal recombination, which concatenates two integers (e.g., “5”, “50”) that appear in the same line and have none but a point “.” or “，” between them. Figure 4(d) and 4(e) show the resegmented results of Figure 4(b) and 4(c) respectively, where the decimal numbers were correctly recombined (except for a misrecognized QXPEHU in Figure 4(e)).

After the resegmentation, we synthesized the results of the Chinese OCR engine and the English OCR engine to achieve an optimal recognition result. We first initiated a synthesis document by simply copying the result recognized by the Chinese engine. For each word $w_{cn}$ with a bounding box $B_{cn}$ in the document, we found the word $w_{en}$ recognized by the English engine whose bounding box $B_{en}$ has the largest overlap with $B_{cn}$. Then we replaced $w_{cn}$ with $w_{en}$ in the document if $w_{en}$ and $w_{cn}$ fulfill the following three criteria: 1) $w_{cn}$ has at least one digit character; 2) the confidence value of $w_{en}$ is not less than that of $w_{cn}$; and 3) the string length of $w_{en}$ is not shorter than that of $w_{cn}$. Using this rule, we adopted the correct numerical values recognized by the English engine while avoiding its garbled Chinese text. Figure 4(f) shows a synthesized result of 4(d) and 4(e), where the wrong numbers of the two engines were corrected.

From the recognized text, we then annotated relevant medical data. Multiple types of medical records were involved, and in general different categories of data entities can be identified in different types of documents. For example, medication names can only be detected in prescriptions and lab test items can only be identified in lab test reports. Therefore, a document classifier was first built to determine the document type. Besides, since the layouts of Chinese medical records are not standardized and can be very flexible, the location-based methods cannot be used to annotate entities of interest and build entity correlations. Thus, we applied flexible entity annotation algorithms such as fuzzy term matching and regular expression matching to locate and identify the entities of interest, and used a slot-based correlation annotation algorithm to locate the related entities (e.g., pairs of test item names and test item values).

1) Document type annotation: keyword-based approach and machine learning approach

Document type annotation is a general problem in the document classification/categorization area. In this study, we focused on two document types, prescription and lab test report. And in each type of documents, there indeed exist some typical keywords which are able to differentiate each other, therefore a keyword-based document classification approach was used. Having explored hundreds of prescriptions and lab test reports, we collected typical keywords for them. For example, the keywords for prescription include {处方(prescription), 药房(pharmacy), 资费(medical fee), 发药(drug dispensing), …}; and those for lab test report include {检验(test), 标本(sampl specimen), …}. Each keyword set was then written as one regex (regular expression) pattern to facilitate the matching process. Once a medical records is matched against a regex pattern, the document type is determined accordingly.
However, since medical records are multifarious and OCR is error-prone, for some documents, neither of the keywords can be matched. In these cases, we also applied the Naïve Bayes model, which is a machine learning model commonly used for document classification, to annotate the documents. The Naïve Bayes approach constructs a classifier by calculating the probability distribution of selected features from training dataset, and then assigns document labels to problem instances. In this study, the training dataset came from an EHR system in China, which contains 2319 prescriptions and 2803 lab test reports. We used bag-of-words as features, and took “tf-idf” (term frequency and inverse document frequency) instead of “tf” (term frequency) to minimize the noise of common words.

2) Entity annotation: dictionary-based approach and regex-based approach

Named entity recognition (NER) is a subtask in information retrieval, which seeks to locate and classify elements in text into pre-defined categories, such as medical terms (e.g., diagnosis names, medication names and test item names) and numerical values, using linguistic-based approaches and statistical models. Differing from general-purposed NER approaches, there is little context information in medical records; instead, hint information like entity inner structures and closed sets of medical terminologies can be used to develop specific annotation algorithms.

For medical term entities, a dictionary-based approach was used to detect the matches of dictionary terms in text, which required a dictionary with high coverage and good quality. By collecting the standard medical terminologies like ICD-10 and the local medical terminology systems in China, we developed a dictionary-based engine to annotate medical terms. Considering the non-standard terminology usage and inevitable OCR errors, a fuzzy matching algorithm was used by calculating the similarity between a recognized word and a dictionary term:

\[
similarity = w_1 \times \text{unigram similarity} + w_2 \times \text{bigram similarity},
\]

where \( w_1, w_2 \in [0, 1] \), and \( w_1 + w_2 = 1 \). Unigram similarity is the ratio of intersected single words to the total number of single words, and bigram similarity is the ratio of intersected two consecutive words to the total number of two consecutive words. If the similarity is above a pre-defined threshold, the term is taken as an entity candidate. Multiple entity candidates for one word are allowed, which can be finally determined during correlation annotation.

Besides, a regex-based approach was used to annotate the entities with obvious structural patterns, such as numerical value, date and identifier. For example, a numerical value can be identified using the regex pattern “[0-9]+(\.[0-9]+)?”.

3) Correlation annotation: slot-based approach

Correlation annotation aims at locating the relationship among entities (e.g., the correspondence between test item names and test item values). Correlation annotation is usually solved using syntax parser or semantic parser within a given context. Since little syntax information is available in printed medical records, we adopted the semantic approach and leveraged layout and entity type information, where the correlations were represented in pre-defined slot-and-filler formats. A pivot entity (e.g., test item name) must be first identified from the entity annotation results (i.e., entity candidates), then the correlated entity (e.g., test item value) of the pivot can be located according to the entity type. Since there are usually multiple correlations of a same type, to avoid incorrect correlation annotation, two entity-span constraints were applied: 1) the correlated entities must occur in a same line or in two consecutive lines; 2) the correlated entities cannot be interrupted by another pivot entity with the same type.

PHR Construction

Before the annotated entities were adopted to construct PHR documents, we checked whether each entity is credible and rejected the entities with low confidence. Both the confidence value \( v \) that represents the OCR engine’s certainty and the similarity value \( s \) were considered. For a medical term entity, \( s \) is the similarity between the word recognized by the Chinese OCR engine and the annotated dictionary term. And for a numerical value, \( s \) is the similarity between the words recognized by the two OCR engines. If the product \( v \times s \) is not less than a pre-defined confidence threshold \( \theta \), then this entity is accepted. Otherwise, it is rejected and not be adopted in building PHR documents.

Finally, we built structured PHR documents from the accepted entities. For each document type such as prescription and lab test report, a template PHR document was defined, as well as mappings from correlated entities to elements in the template document. The template documents are based on the interface XML format used in our EHR system, which can be easily transformed into the standard CDA (HL7 Clinical Document Architecture) format. Given the annotation results of a document, a template PHR document is first determined according to the annotated document type. Then the accepted entities are mapped to the elements in the PHR document. As shown in the example in Figure 2, three pairs of correlated entities (“Name” and “Value” in the “Lab Test” correlations) were transformed to three elements (“TestResultObservation”) in a PHR document of lab test report.
Results

To evaluate the performance of our methods, we validated the PHR documents automatically built by our approach against the gold standard created by human readers using the corpus described above which has 100 testing documents. The exact match evaluation was performed, where an automatically extracted entity is correct if its text exactly matches that of the gold standard entity (i.e., “5.5” and “5.50” are not regarded as equal). For each experiment configuration, precision (positive predictive value), recall (sensitivity) and F-measure (harmonic mean of precision and recall) were computed.

In our study, we first configured our system pipeline as described in the above section, and performed the evaluation for the confidence threshold value $\theta = 0.0$ (i.e., all the recognized and annotated entities are adopted to generate PHR documents). Table 1 demonstrates the results for every entity category and for all entities. Here the category of “Term” means the union set of the categories of diagnosis name, medication name and test item name. As shown in Table 1, without constraining the confidence value, our system could identify over 80% (recall) entities that were annotated by human readers, and over 90% (precision) of our extracted entities were precise. The performance for the term entities, with a recall of 0.88, outweighed that for the value entities (i.e., test item values), with a recall of about 0.75.

Table 1. Evaluation on the system using the confidence threshold = 0.0 (P: precision; R: recall; F: F-measure)

<table>
<thead>
<tr>
<th>Category</th>
<th>#entity</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis name</td>
<td>84</td>
<td>0.973</td>
<td>0.869</td>
<td>0.918</td>
</tr>
<tr>
<td>Medication name</td>
<td>111</td>
<td>0.916</td>
<td>0.784</td>
<td>0.845</td>
</tr>
<tr>
<td>Test item name</td>
<td>650</td>
<td>0.946</td>
<td>0.898</td>
<td>0.922</td>
</tr>
<tr>
<td>Test item value</td>
<td>631</td>
<td>0.867</td>
<td>0.746</td>
<td>0.802</td>
</tr>
<tr>
<td>Term</td>
<td>845</td>
<td>0.945</td>
<td>0.880</td>
<td>0.912</td>
</tr>
<tr>
<td>All</td>
<td>1476</td>
<td>0.914</td>
<td>0.823</td>
<td>0.866</td>
</tr>
</tbody>
</table>

To demonstrate the influence of the confidence threshold, we kept the same configuration as above but gradually changed the threshold $\theta$ from 0.0 to 1.0. As shown in Figure 5, with the increasing of the threshold, the precisions of the system slowly rose and could achieve over 0.98 for the term entities and 0.94 for the value entities, whereas the recalls and the F-measures dramatically descended accordingly.

![Figure 5. Precision, recall and F-measure of the system with respect to the confidence threshold](image)

Besides, we also performed the evaluation on different pre-processing and post-processing algorithms for our system. Table 2 shows the results of the PHR document evaluation on image pre-processing approaches, where the configurations of the other modules except pre-processing were set in the same manner as the experiment shown in Table 1. Compared with the results of directly using original images, the image processing algorithms we used did not significantly improve the precision of the system. However, all the three algorithms promoted the recall to some extent, and the combination of these algorithms achieved the best performance.

Table 2. Evaluation results for pre-processing

<table>
<thead>
<tr>
<th>Pre-processing algorithm</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.892</td>
<td>0.736</td>
<td>0.807</td>
</tr>
<tr>
<td>Binarization</td>
<td>0.911</td>
<td>0.783</td>
<td>0.842</td>
</tr>
<tr>
<td>Denoising + Binarization</td>
<td>0.910</td>
<td>0.812</td>
<td>0.858</td>
</tr>
<tr>
<td>Denois. + Binar. + Deskew.</td>
<td>0.914</td>
<td>0.823</td>
<td>0.866</td>
</tr>
</tbody>
</table>

Table 3. Evaluation results for post-processing

<table>
<thead>
<tr>
<th>Post-processing algorithm</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.760</td>
<td>0.629</td>
<td>0.689</td>
</tr>
<tr>
<td>Resegmentation</td>
<td>0.874</td>
<td>0.781</td>
<td>0.825</td>
</tr>
<tr>
<td>Multi-engine synthesis</td>
<td>0.818</td>
<td>0.671</td>
<td>0.737</td>
</tr>
<tr>
<td>Reseg. + Synthesis</td>
<td>0.914</td>
<td>0.823</td>
<td>0.866</td>
</tr>
</tbody>
</table>
Table 3 shows the results of the PHR document evaluation on different configurations of post-processing approaches. Both the word resegmentation algorithm and the multi-engine synthesis algorithm obviously improved the precision and recall of the system, and their combination obtained the best results.

To evaluate the performance of the document type annotation algorithms, we directly compared the types classified by the algorithms with the type tags given by the human readers for the 100 testing documents, and computed the classification accuracy. As shown in Table 4, both the keyword-based and Naïve Bayes algorithms led to a few classification errors, and the combination of these algorithms achieved an absolutely correct result on our testing set.

To evaluate the performance of the term matching approaches, we performed the PHR document evaluation on the exact matching algorithm as well as the fuzzy matching algorithms using unigram similarity and using the combination of unigram and bigram as Equation (1) \((w_1 = 0.4, w_2 = 0.6)\). As shown in Table 5, though the precision of the exact matching algorithm was the highest, its recall was relatively low. The recalls of the two fuzzy matching metrics had no significant difference, but the precision of the combined metric was obviously higher than that of unigram.

To evaluate the applicability of the proposed system, we also performed an experiment on same medical record documents photographed by different users using different mobile phones. 20 printed documents were randomly selected from the corpus and 5 volunteer users respectively shot the 20 documents using their own mobile phones. Figure 6 shows examples of a same document photographed by different users, where the clarity and illumination distribution of the images are obviously different. The quality of the photos shot by user #5 is on the low side in comparison to others. We performed our pipeline on these 20 documents for the 5 users respectively, using the same configuration as the experiment shown in Table 1, and the evaluation results are shown in Table 6. For the images from each user, our system achieved a precision of over 0.85. And the recalls for the images from all the users except #5 were also close (from 0.75 – 0.79).

To analyze the extraction errors of our system based on the results of the experiment shown in Table 1. We detected total 284 errors including 169 missing entities that were annotated by the human readers but not identified by the system, 23 superfluous entities that were not manually annotated but extracted by the system, and 92 incorrect entities of which the results recognized by the system were different with the gold standard. The reasons of the errors were also classified by the human readers, which are shown in Table 7. More than a half of the errors were caused by the OCR engines, and the annotation algorithms also contributed to a large proportion of errors.
Admittedly, the image quality of photographed medical records significantly affects the performance of OCR and annotation. There are two primary image problems that may cause extraction errors: 1) printing and typesetting defects in original printed documents; 2) problems caused by camera or careless photographing. As shown in the experimental results, a majority of the image problems could be solved by the image pre-processing approach and the fuzzy term matching algorithm. The overall system had good tolerance to uneven light and shade, Gaussian noise, and skewing in images, and was adaptable to photos shot by different users using different phones. However, as shown in Table 7, there were still some extraction errors caused by image defects that were not eliminated. Figure 7(a) gives some examples of printing defects where some strokes were not completely printed, and Figure 7(b) shows a typesetting problem where two entities (test item name and value) are concatenated with each other. Besides, more errors tended to occur when the system handled images with low clarity (e.g., some images shot by user #5).

As shown in Table 2, the image pre-processing approach significantly improved the sensitivity of our system, because the uneven illumination distribution, Gaussian noise and skewing problems could be relieved. However, as shown in Figure 7(c), these image processing algorithms also made a few additional errors, where some extremely thin strokes disappeared in the binarized images. Even so, we still applied these algorithms in the current system because the overall performance could be promoted. And a detail-preserving image processing approach can be used to solve this problem in the future.

The limitation of the OCR engines caused more than a half of the extraction errors. Although we did not attempt to modify the algorithms of the OCR engines, we observed that reasonable post-processing algorithms, such as word resegmentation, could greatly improve the overall performance (see Table 3). More importantly, we observed that with well-defined synthesis criteria, the results of multiple OCR engines could be combined to correct some errors from each OCR engine alone. In this study, we only synthesized the results of one Chinese OCR engine and one English OCR engine. It can be predicted that if more OCR engines are integrated in a reasonable manner, OCR errors can be further reduced. As shown in Table 7, the post-processing algorithms produced very few additional errors (one of which has been shown in Figure 4(f)), and the benefits significantly outweigh the drawbacks.

Another key factor that contributed to the extraction errors is the annotation algorithms. We observed two main types of annotation errors. The first type was the superfluous term entities caused by the fuzzy term matching algorithm. Compared with the exact matching algorithm, the fuzzy similarity metric greatly improved the recall of term identification, but indeed incorrectly matched some entities that were not regarded by human readers as terms of interest, and therefore reduced the precision as shown in Table 5. Since the performance goal of the system was not only precision but also recall, we finally used the fuzzy matching algorithm in the system. Another type of errors was the item values that were exactly recognized by the OCR engines and correctly annotated as numerical value entities, but were not correctly linked to their corresponding item names by the correlation annotation algorithm. That is primarily because the location information of the entities is not adequately used by the slot-base algorithm. For the value entities that were correctly identified, the current slot-based algorithm could achieve an approximately 86% recall of correlation annotation, which can probably be further improved by combining statistical learning models such as conditional random field. Besides, though we achieved a 100% accuracy of document classification on our testing set by combining the keyword-based and Naive Bayes approaches, more testing data in the document type level is still needed to further prove the effectiveness of the classification algorithm.

Although the confidence threshold of the system can be configured, we did not finally use it in the current system configuration (i.e.,  $\theta = 0.0$). That is because the performance goal was to achieve both high precision and high recall, and we observed that with the increasing of the threshold, the recall descended more dramatically than the precision increased (see Figure 5). However, if a case requires a higher precision, we can accordingly raise the threshold (and can also use the exact term matching algorithm instead of the fuzzy matching algorithm as discussed above).
Conclusion

Data collection, which is the basis of building PHRs, is very difficult in a developing country like China due to the policy and technical obstruction to giving patients access to EHR data. Considering the easy availability of printed medical records and the high popularity of smartphones, we proposed a practical approach to build structured PHRs from printed medical records photographed by mobile phones. By combining a series of image processing, OCR, pattern matching and machine learning techniques in a reasonable system pipeline, we addressed the problems of low image quality, layout diversity and content complexity of photographed medical records. The proposed approach was applied to build PHRs from real world prescriptions and lab test reports, and the results showed that our approach can automatically extract structured medical data with relatively high precision and recall, and has wide applicability.

As discussed above, the limitations of the current approach caused some errors that could be eliminated by improving the architecture and algorithms in the future, including detail-preserving image processing, synthesis of more OCR engines, as well as layout-aware and statistic-based annotation. This paper focused on the extraction of medical information, such as diagnosis, medication and test item, from medical records. Actually, the privacy information specified by HIPAA, such as patient information and dates, can also be located using our methodology for entity and correlation annotation, which can then be de-identified for data transfer and distribution purpose. Besides, though the current system was designed for printed Chinese medical records, our methodology can also be adapted to those of other languages or hand-written medical records by extending the OCR and annotation modules.

References

5. Kmietowicz Z. Patients will have digital access to GP records by 2015, says NHS England. BMJ. 2014 Nov 12;349:g6805.

Yanhua Lin, PhD1, Catherine J Staes, MPH, PhD1, David E Shields1, Vijay Kandula, MD1, Brandon M Welch, MS, PhD2, Kensaku Kawamoto, MD, PhD, MHS1

1Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 2Biomedical Informatics Center, Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC

Abstract
When coupled with a common information model, a common terminology for clinical decision support (CDS) and electronic clinical quality measurement (eCQM) could greatly facilitate the distributed development and sharing of CDS and eCQM knowledge resources. To enable such scalable knowledge authoring and sharing, we systematically developed an extensible and standards-based terminology for CDS and eCQM in the context of the HL7 Virtual Medical Record (vMR) information model. The development of this terminology entailed three steps: (1) systematic, physician-curated concept identification from sources such as the Health Information Technology Standards Panel (HITSP) and the SNOMED-CT CORE problem list; (2) concept de-duplication leveraging the Unified Medical Language System (UMLS) MetaMap and Metathesaurus; and (3) systematic concept naming using standard terminologies and heuristic algorithms. This process generated 3,046 concepts spanning 68 domains. Evaluation against representative CDS and eCQM resources revealed approximately 50-70% concept coverage, indicating the need for continued expansion of the terminology.

Introduction
Need for data standardization
Despite the demonstrated potential for clinical decision support (CDS) to improve care quality and promote patient safety (1-4), CDS availability continues to be limited in most clinical settings (5-7). An important reason for this limited CDS availability is the difficulty of scaling CDS across institutions (8-10), with the lack of data standardization being a predominant barrier to sharing (11). Electronic clinical quality measurement (eCQM), which shares many requirements with CDS and can be implemented using a common underlying system (12), has a similar need for standardized data. Indeed, the U.S. Office of the National Coordinator for Health IT and the Centers for Medicare & Medicaid Services are sponsoring an initiative known as the Clinical Quality Framework to develop a harmonized set of standards to fulfill the needs of both CDS and eCQM (13).

Figure 1 provides an overview of aspects of data standardization for CDS and eCQM. One aspect of standardization is the information model, which identifies data classes (e.g., Problem), attributes (e.g., problem code), and the relationship of classes to one another (e.g., the relationship of Problems to Encounters; not shown). Coded attributes
describe concepts such as “diabetes mellitus,” which in turn may be defined by a value set of instance codes that are indicative of the concept (e.g., SNOMED-CT 314902007, type II diabetes mellitus with peripheral angiopathy).

**Need for a concept terminology for CDS and eCQM**

Data standardization efforts in CDS and eCQM have generally focused on standardization of (i) the information model, (ii) the superset of instance codes that may be used within coded attributes, and, in some cases, (iii) individual value sets (14, 15) However, to our knowledge, there has been no systematic effort to define a common concept terminology for CDS and eCQM to facilitate knowledge sharing and semantic interoperability.

Many standard terminologies, such as SNOMED-CT, RxNorm and LOINC, are available for use in CDS and eCQM with relatively adequate breadth, depth, and granularity (16). However, the sheer volume of concepts in these terminologies can make it challenging to ensure that different CDS and eCQM implementers choose the same concepts in their respective implementations. For example, the number of coded concepts in the UMLS Metathesaurus (>1,400,000), SNOMED-CT (> 310,000), RxNorm (> 93,000), LOINC (> 46,000), and ICD-10 (> 12,000) makes consistent concept selection challenging (17). Meanwhile, many terminologies remain semantically incompatible (18). The diversity in different terminological systems hampers the possibility of sharing and reasoning with data within different systems (11). Therefore, the challenge lies less with the lack of relevant standards, but more with the fact that multiple terminologies are in concurrent use (18), and with the sheer volume of concepts. Furthermore, the lack of hierarchical structures in some terminologies makes it difficult to find useful terms that are less specific, as is often needed for CDS and eCQM. Consequently, it is imperative to identify and maintain a much smaller subset of broader concepts with utility for computerized CDS and eCQM. Here, we describe an effort to meet this need within the context of OpenCDS, which is a multi-institutional collaborative initiative to develop open-source, standards-based tools and resources to enable CDS and eCQM at scale (12, 19).

**Methods**

**Project context and operational use of terminology**

The concept terminology was developed in the context of the OpenCDS effort to support CDS and eCQM. OpenCDS has been implemented in a number of electronic health record (EHR) systems and provides a reference implementation of the HL7 vMR data model standard (12, 19-22). The vMR was designed originally for CDS but has been subsequently applied to eCQM as well (12). The vMR contains 68 coded attributes, such as Adverse Event, Encounter Type, Goal Focus, Observation Focus, Problem, Procedure, Medication, and Supply (Table 1).

![Figure 2. Use of concept terminology in OpenCDS knowledge authoring.](image)

In OpenCDS, CDS or eCQM modules are authored as a series of human-readable rules and then translated into machine-executable knowledge. Concepts are accessed via drop-down lists specific to the type of concept involved (e.g., gender) (Fig. 2). These concepts, in turn, are mapped to value sets containing applicable local or standard codes. The use of concepts enables a clear separation of concerns between terminology mapping and logic authoring.
### Table 1. Summary of coded attributes in the vMR information model

<table>
<thead>
<tr>
<th>Entities: Terms Relating to All Entities</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Type</td>
<td>Observation Coded Value</td>
</tr>
<tr>
<td>Entity Relationship</td>
<td>Observation Criticality</td>
</tr>
<tr>
<td>Person</td>
<td>Observation Focus</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Observation Interpretation</td>
</tr>
<tr>
<td>Gender</td>
<td>Observation Method</td>
</tr>
<tr>
<td>Preferred Language</td>
<td>Observation Target Body Site</td>
</tr>
<tr>
<td>Race</td>
<td>Observation Target Body Site Laterality</td>
</tr>
<tr>
<td>Substance</td>
<td>Observation Unconducted Reason</td>
</tr>
<tr>
<td>Manufacturer</td>
<td>Problem</td>
</tr>
<tr>
<td>Medication</td>
<td>Problem Affect Body Site</td>
</tr>
<tr>
<td>Medication Branded</td>
<td>Problem Affect Body Site Laterality</td>
</tr>
<tr>
<td>Medication Generic</td>
<td>Problem Importance</td>
</tr>
<tr>
<td>Substance Form</td>
<td>Problem Severity</td>
</tr>
<tr>
<td>Supply</td>
<td>Problem Status</td>
</tr>
<tr>
<td>Supply</td>
<td>Procedure</td>
</tr>
<tr>
<td></td>
<td>Procedure Approach Body Site</td>
</tr>
<tr>
<td></td>
<td>Procedure Approach Body Site Laterality</td>
</tr>
<tr>
<td></td>
<td>Procedure Criticality</td>
</tr>
<tr>
<td></td>
<td>Procedure Method</td>
</tr>
<tr>
<td></td>
<td>Procedure Target Body Site</td>
</tr>
<tr>
<td></td>
<td>Procedure Target Body Site Laterality</td>
</tr>
<tr>
<td></td>
<td>Unconducted Procedure Reason</td>
</tr>
<tr>
<td>Clinical Statements: Terms Relating to All Clinical Statements</td>
<td>Substance Administration</td>
</tr>
<tr>
<td>Clinical Statement Relationship</td>
<td>Dose Type</td>
</tr>
<tr>
<td>Data Source Type</td>
<td>Dosing SIG</td>
</tr>
<tr>
<td>Adverse Event</td>
<td>Information Attestation Type</td>
</tr>
<tr>
<td>Adverse Event Affect Body Site</td>
<td>Substance Administration Approach Body Site</td>
</tr>
<tr>
<td>Adverse Event Affect Body Site Laterality</td>
<td>Substance Administration Approach Body Site Laterality</td>
</tr>
<tr>
<td>Adverse Event Agent</td>
<td>Substance Administration Criticality</td>
</tr>
<tr>
<td>Adverse Event Criticality</td>
<td>Substance Administration General Purpose</td>
</tr>
<tr>
<td>Adverse Event Severity</td>
<td>Substance Administration Target Body Site</td>
</tr>
<tr>
<td>Adverse Event Status</td>
<td>Substance Administration Target Body Site Laterality</td>
</tr>
<tr>
<td></td>
<td>Substance Delivery Method</td>
</tr>
<tr>
<td></td>
<td>Substance Delivery Route</td>
</tr>
<tr>
<td></td>
<td>Undelivered Substance Reason</td>
</tr>
<tr>
<td>Encounter</td>
<td>Supply</td>
</tr>
<tr>
<td>Encounter Type</td>
<td>Supply Criticality</td>
</tr>
<tr>
<td>Encounter Criticality</td>
<td>Supply Target Body Site</td>
</tr>
<tr>
<td>Appointment Proposal Criticality</td>
<td>Supply Target Body Site Laterality</td>
</tr>
<tr>
<td>Appointment Request Criticality</td>
<td>Undelivered Supply Reason</td>
</tr>
<tr>
<td>Goal</td>
<td>Goal Criticality</td>
</tr>
<tr>
<td>Goal Focus</td>
<td>Goal Focus</td>
</tr>
<tr>
<td>Goal Status</td>
<td>Goal Status</td>
</tr>
<tr>
<td>Goal Target Body Site</td>
<td>Goal Target Body Site</td>
</tr>
<tr>
<td>Goal Target Body Site Laterality</td>
<td>Goal Target Body Site Laterality</td>
</tr>
<tr>
<td></td>
<td>Undelivered Substance Reason</td>
</tr>
</tbody>
</table>

**Objectives and requirements**

The objectives of this project were to (i) define a standard and extensible approach for curating a concept terminology for CDS and eCQM that can be leveraged by the OpenCDS community and to (ii) populate the terminology with an initial set of common, high-level concepts useful for CDS and eCQM knowledge authoring. Requirements included (i) adherence to the 80-20 rule, with a goal of initial inclusion of high coverage of concepts likely to be needed for
typical CDS or eCQM use cases; (ii) the leveraging of standard terminologies; (iii) internal consistency; and (iv) avoidance of duplicate concepts.

**Overview of approach**

The terminology was developed using three steps. First, relevant concepts were identified in a systematic, physician-curated manner. Second, concepts were de-duplicated using UMLS MetaMap and Metathesaurus. Finally, concepts were named using standard terminologies and heuristic algorithms. These steps are outlined in greater detail below.

**Step 1a: Candidate concept identification**

To identify relevant concepts, we first reviewed the Healthcare Information Technology Standards Panel (HITSP)’s Clinical Document and Message Terminology specification (HITSP C80, Version 2.0.1) (23). This document defines the vocabularies used by HITSP specifications for clinical documents and messages to support the interoperable transmission of information. If this specification defined a finite value set for a targeted coded attribute type, that value set was used as the set of candidate concepts for physician review and curation in the next phase of this step.

If HITSP C80 did not define a value set for a coded attribute type, we next searched the Public Health Information Network Vocabulary Access and Distribution System (PHIN VADS) and Value Set Authority Center (VSAC). If an appropriate value set was identified here, then that value set was identified as the candidate concept set.

If the above resources did not identify a relevant value set, or if the value set identified was extremely large in scope, the potential concepts for physician review were restricted using various methods. For example, HITSP C80 recommends concepts using the Veteran Administration and Kaiser Permanente (VA/KP) problem list subset of SNOMED-CT for describing problems (23). However, the VA/KP problem list subset contains over 15,000 concepts, making it a challenge to review. Therefore, we instead used the Clinical Observations Recording and Encoding (CORE) subset of SNOMED-CT (24). CORE was based on datasets submitted by 8 institutions, and it is a frequency-based approach to problem list development. Compared to VA/KP, CORE is smaller, and 94.8% of coded problem entries from Brigham and Women's Hospital are in the CORE subset (4), indicating high coverage of used concepts. For our purposes, we started with the 266 CORE problem list entries that were reported by most (8 or 7) of the institutions as the candidate set of problem concepts for potential inclusion in the initial CDS/eCQM terminology.

Similar methods were used for enriching the candidate set of concepts designated for physician review. For example, laboratory test concepts were restricted to the LOINC Universal Laboratory Order Codes, whose approximately 300 codes cover more than 95% of the lab test orders in the United States (25).

**Step 1b: Physician curation**

After candidate concepts were identified in the step above, a physician informaticist (VK) who is a practicing hospitalist reviewed each concept in the candidate set and identified those that have a reasonable likelihood of being useful for CDS purposes based on personal experience. The physician informaticist classified these concepts into 4 categories: 1 - high priority, 2 - moderate priority, 3 - low priority, and 4 - not appropriate. Concepts with priority 1 and 2 were uploaded into the Apelon Distributed Terminology System (DTS) terminology server for management.

**Step 2: De-duplication**

Duplicate entries are a common problem in terminologies (6, 26), even in the UMLS Metathesaurus (27). To identify and deprecate duplicate concepts, we implemented a systematic methodology for identifying potential duplicates, which were verified through physician review (Fig. 3).

Before starting, we identified candidate concepts for de-duplication by excluding concepts that had previously been deprecated or were being used for administrative purposes (e.g., to name a specific quality measure, such as HEDIS Breast Cancer Screening). We then searched for exact string matches to SNOMED-CT terms (including synonyms) in the UMLS Metathesaurus, capturing the corresponding UMLS concept unique identifier (CUI) through the process. This subset of the terminology (T1) represented a large portion of the original set, indicating that SNOMED-CT was a reasonable source for concept names. The remaining OpenCDS concepts (R1) were then screened for perfect string matches to the terms in any other standard terminologies in UMLS, and the results were processed similarly (T2).

For the remaining concepts with no perfect string matches available (R2), the UMLS MetaMap tool (28) was used to identify potential matching UMLS CUIs. Concepts that could not be matched to any UMLS terms in this manner
(R3) were generally unique concepts used as intermediate conclusions (e.g., “age $\geq 50$ and $< 75$ years”) and were not processed further for de-duplication.

Next, we combined all the concepts and the corresponding UMLS CUIs from T1, T2 and T3 and identified potential duplicate concepts sharing the same CUIs. These potential duplicates were reviewed by a physician. If two or more concepts shared the same CUI but were deemed to be distinct, we updated the CUI for one of the concepts using the UMLS Metathesaurus. If two or more concepts were deemed to be duplicative, one was kept and the rest were deprecated.

![Strategy to identify duplicate concepts](image)

**Figure 3.** Strategy to identify duplicate concepts.

**Step 3: Concept naming**

For concepts matched to more than one CUI, the preferred term for each CUI was obtained from the UMLS Metathesaurus and reviewed to identify the most appropriate CUI for the concept. Next, using the CUI associated with each concept, preferred terms from appropriate standard terminologies were obtained by leveraging the UMLS as shown in Fig. 4. Finally, concept names were post-processed for consistency using heuristic algorithms. For example, capitalization schemes were standardized. Also, concepts were named as the first major category followed by any modifiers to facilitate finding all variations on a root concept in a drop-down list. For example, “Bilateral Mastectomy” was renamed “Mastectomy, Bilateral” and “Lower Extremity Amputation” was renamed “Amputation, Lower Extremity.”
Evaluation of concept coverage
We evaluated the degree of coverage of the concept terminology for sample CDS and eCQM knowledge resources. For CDS, we reviewed the data sections of example Arden Syntax Medical Logical Modules provided in the appendix of version 2.8 of the standard (29). For eCQM, we reviewed the first 50 value sets in the National Quality Forum’s eCQMs from 2011 (30).

Results
Concept identification, de-duplication, and naming
3886 concepts spanning the 68 vMR coded attributes were identified for potential inclusion in the terminology. Following physician informaticist review, approximately 2,200 clinical concepts were selected for inclusion.

![Figure 4](image)

Figure 4. Strategy to identify standard terms for local terminology.

Our systematic de-duplication method identified 110 potential duplicates. After review by a physician, 72 concepts were confirmed to be duplicates and deprecated. For example, “Urinary retention” and “Retention of urine” were found to be duplicates, leading to one of the concepts being deprecated. Finally, 1,928 concepts with UMLS CUIs were named using SNOMED-CT, RxNorm, and other standard terminologies included in the UMLS.

Concept upload to Apelon DTS terminology server
Concepts were then uploaded into the Apelon DTS terminology server and mapped to corresponding coded attribute types. Each of these concepts from the external terminologies then became a unique Apelon DTS concept, and a code was assigned automatically to the concept. All concept names were capitalized (proper case). This import also updated the hierarchical relationships so that concepts are the descendants of corresponding vMR coded attribute types. In some cases, a single concept can be associated with two or more coded attribute types, but concepts were defined only once. For example, “Pregnancy Test” can be either an Observation Focus with a possible result value or simply a Procedure that was performed. Accessing “Pregnancy Test” from Observation Focus or Procedure in OpenCDS will bring the user the same concept and code (C1693).

Concept coverage
The terminology created was evaluated against a previously unseen set of Arden Syntax Medical Logic Modules and National Quality Forum eCQMs. This analysis showed that the terminology developed covered approximately 70% of the concepts referenced in the Medical Logic Modules and approximately 50% of the concepts referenced in the
eCQMs. Many of the concepts that were not covered by the terminology consisted of concepts for specific medications.

Discussion
Implementing CDS capabilities usually requires several terminologies due to their different domain(s) of coverage and granularity (20, 31). As a result, concurrent use of different terminologies is a significant challenge, and a CDS resource designed for use in one setting may not be readily usable in another setting that uses a different set of terminologies, even when similar concepts are being captured (18). Furthermore, when different institutions use different subsets with non-overlapping terms, significant interoperability challenges occur (4). To address these issues, we built a terminology for CDS and eCQM based on the HL7 vMR information model as a part of the OpenCDS initiative. A central part of this terminology development effort was the definition and application of systematic approaches to de-duplication and naming standardization.

Achieving semantic interoperability for CDS and eCQM depends on the use of common information models and common associated concepts (32). In our study, we sought to define a “starter set” of concepts that have a reasonable likelihood of being useful for CDS or eCQM purposes. However, identifying what terminology is “best” or which term is “common” is challenging. For example, some concepts that are common to ambulatory care may not be relevant in an inpatient scenario. Thus, besides the domain-specific expertise from our group, our strategy was to leverage established sources of common concepts such as the SNOMED-CT CORE problem list and HITSP recommendations.

Implementing effective vocabulary control in medical informatics includes facilitating the selection of the most appropriate concept for a given clinical scenario (33). Wright et al. used human immunodeficiency virus (HIV) as an example to demonstrate the importance of screening the concepts for a specific data attribute. In this example, they note that SNOMED-CT has 138 concepts related to HIV and that, without appropriate filtering, a clinician may easily select an incorrect code by mistake (4). We believe our terminology consisting of common and relevant concepts will lead to less chances for inadvertent selections of inappropriate concepts during knowledge authoring.

We included concepts from standard terminologies, such as SNOMED-CT, LOINC, and HL7. In addition, we have added concepts without correspondence to standard terms, primarily administrative concepts such as intermediate conclusions (e.g., “Denominator Inclusion Criteria Met”) or quality measure specifications (e.g., “HEDIS Frequency of Prenatal Care Measure”). Thus, the OpenCDS terminology brings concepts together from disparate controlled terminologies and non-standard terminologies into a single conceptual dictionary of medical concepts. This approach is supported by the vMR information model, which can make use of both standard and local codes. Although OpenCDS can make use of data expressed in many different medical terminologies, it does so through the use of OpenCDS concepts, which map one or more specific and concrete codes from standard or proprietary medical terminologies to a single OpenCDS concept code. An OpenCDS concept is the interface between the clinical ideas and the data details that represent instantiations of the clinical concepts. The clinical rules use OpenCDS concepts in preference to references to the raw data, and terminology mappings provide implementations of those concepts as value sets of codes from one or more code systems. This separates the logic of the rules from the details of the data that the rules work on, thereby facilitating knowledge authoring, maintenance, and sharing.

We learned some lessons when building and maintaining the foundational terminology. When adding concepts in the future, a careful analysis is needed to determine if a concept closely relates to an existing concept. Care should also be taken to ensure consistent naming schemes (34). To avoid ambiguity and to offer an easy way to identify duplicates during maintenance, full names should be provided, either directly or as a concept property.

Building this terminology is an ongoing task. As identified in the evaluation, while the terminology had substantial coverage of relevant concepts, there were still significant gaps in the content. To address this need for continual enhancement and maintenance, we are developing standard operating procedures for adding new content in a systematic, consistent, and non-duplicative manner. In particular, we are seeking to make it easier for OpenCDS users who are not a part of the core development team to request the addition of new concepts.

While the focus of this study was on CDS and eCQM, there are potentially many other contexts in which a standards-based, domain-optimized terminology could provide value. For example, clinical specialties may wish to define common concepts used in their domains, and information systems may benefit from the definition of common concepts.
used in specific sub-systems (e.g., order entry). While we have not yet applied our methodology to the development of any additional terminologies, we speculate that the methods used in this study—entailing systematic concept identification, deduplication, and naming—could be helpful in developing and maintaining such domain-optimized terminologies.

Conclusions
In this paper, we shared our experiences in building and maintaining a terminology for CDS and eCQM, which is in active use within the OpenCDS community. To the best of our knowledge, this is the first terminology developed specifically to meet CDS and eCQM needs. These concepts and tools are freely available to the open-source community to use and adapt. Because this terminology was built in reference to a standard HL7 clinical information model, our methods and results are likely to be applicable for implementations in other institutions and settings. We believe our experiences described herein will also be informative for others who seek to maintain controlled terminologies in pursuit of semantic interoperability.

Acknowledgments
The project was supported by NLM training grant (T15LM007124) and the University of Utah Knowledge Management and Mobilization initiative. We thank Bruce E. Bray, Phillip B. Warner, Tyler J. Tippettts, Polina V. Kukhareva, and Alisha Edison-Hair for their assistance and fruitful discussion, Ryan Butcher for terminology domain-specific suggestions, and Jack Bowie from Apelon for answering technical questions. KK is currently or recently served as a consultant on CDS to the Office of the National Coordinator for Health IT, ARUP Laboratories, McKesson InterQual, ESAC, Inc., JBS International, Inc., Inflexxion, Inc., Intelligent Automation, Inc., Partners HealthCare, the RAND Corporation, and Mayo Clinic. KK receives royalties for a Duke University-owned CDS technology for infectious disease management known as CustomID that he helped develop. KK was formerly a consultant for Religent, Inc. and a co-owner and consultant for Clinica Software, Inc., both of which provide commercial CDS services, including through use of a CDS technology known as SEBASTIAN that KK developed. KK no longer has a financial relationship with either Religent or Clinica Software. KK has no competing interests related to OpenCDS, which is freely available to the community as an open-source resource. VK is a part owner of Coresys Infotech, a CDS company. All other authors declare no conflict of interest.

References
Analysis of empty responses from electronic resources in infobutton managers

Jie Long, PhD1, Nathan C. Hulse, PhD1,2, Cui Tao, PhD3
1Intermountain Healthcare, Salt Lake City, UT; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 3University of Texas School of Biomedical Informatics, Houston, TX;

Abstract

Infobuttons provide context-aware educational materials to both providers and patients and are becoming an important element in modern electronic health records (EHR) and patient health records (PHR). However, the content from different electronic resources (e-resource) as responses from infobutton manager has not been fully analyzed and evaluated. In this paper, we propose a method for automatically analyzing responses from infobutton manager. A tool is implemented to retrieve and analyze responses from infobutton manager. To test the tool, we extracted and sampled common and uncommon concepts from EHR usage data in Intermountain Healthcare’s enterprise data warehouse. From the output of the tool, we evaluate infobutton performance by multiple categories, including against the most and less common used concepts, grouped by different modules in patient portal, by different e-resources, and by type of access (standardized Health Level Seven (HL7) vs not). Based on the results of our evaluation, we provide suggestions for further enhancements of infobuttons to the current implementation, including suggesting accessing priorities of e-resources and encouraging the use of the HL7 standard.

Introduction

Infobuttons have been becoming an increasingly important element in modern EHRs and PHRs and serve as a notable solution for addressing clinical information needs at the point-of-care. Currently infobuttons are an important part in the HL7 standard and well as in the meaningful use. Infobuttons can be used to deliver educational materials to both physicians and patients. Infobuttons provide services to deliver context-aware information, which has proved to be an effective way for retrieving more accurate information to the end users. The context of an infobutton request includes basic information from the user, the patient, the system, and the requested concept. As a result, besides the requested concepts that a user would normally type into the search box, the infobutton manager is intelligent enough to retrieve information from different e-resources and find the best match of an input concept given the provided context. In that way, content from the infobutton response is tailored to be more accurate to reflect the needs of the end users.

It is very important that infobuttons perform well to satisfy users’ information needs, especially in PHRs where the ends users have less professional knowledge than physicians in seeking out good clinical information. However, the content brought up by infobutton manager has not been fully analyzed and its performance is not well-documented in the medical informatics literature. The lack of a full systematic analysis of the content might put designer’s bias into the infobutton manager and result in less than optimal responses. These can result in a frustrating end-user experience, one that could lessen their likelihood of using the tool in the future.

One of the interesting limitations of most current infobutton implementations is that the infobutton manager is typically not aware of the content it will receive before it generates a link that directs users to it. This occurs because the first version of the HL7 standard (the one that is most widely adopted) is a unidirectional standard that does not allow for metadata to flow back to the infobutton manager. As such, there is an implicit trust between the infobutton manager (on the EHR side) and the content provider that there will be relevant content returned. Clearly, content providers have a finite set of available content, and it may not be indexed in such a way that all of it is readily accessible from the infobutton manager. This creates the potential for ‘empty responses’ from e-resources linked to from the infobutton manager, or scenarios in which a user clicks on a link from the infobutton and arrives at a page with zero results. This scenario is to be anticipated at least some of the time when linking from very uncommon concepts, but it becomes notably worse if the infobutton manager returns empty responses in commonly used concepts or highly-trafficked modules.

However, systematically analyzing content brought up by infobuttons is a new challenge. In the current implementation of infobutton managers, the rank of e-resources is predefined based on the designer’s interest. There is nowhere in the current design to collect user’s feedbacks on the content of e-resources from infobutton manager.
In this research, we propose a method to automatically detect empty responses from e-resources and provide statistics on the empty responses. The content from infobuttons is retrieved from Intermountain Healthcare’s patient portal MyHealth (12). The content coverage presented in this manuscript includes three modules in MyHealth and seven different e-resources in these modules. We implemented a tool to automatically detect empty responses for sampled concepts. Each concept sends an infobutton request. These concepts are sampled from the three most trafficked ‘infobutton-supported’ modules in our personal health record, including health concerns (problem list), medications, and tests and procedures (lab values). We implemented a tool that is capable of firing a series infobutton requests (from a list of input parameters), automatically detecting empty responses, and recording data for further analysis. The number of empty responses is then classified by module and by e-resources.

**Hypothesis**

In approaching this evaluation, we anticipate that (1) websites with HL7 standardized API will return fewer empty responses; (2) infobuttons will return fewer empty responses in more commonly used concepts; (3) the most used infobutton module will return the fewest empty responses (4) websites will exhibit variance in supporting different modules and in supporting common and uncommon concepts. We expect that the conclusions help researchers to better understand the vacancy rate in content brought by infobutton and therefore improve the current performance of infobutton. From the distribution of the empty responses, the conclusion also help designers to better assign accessing priorities for different websites so the end user will see less empty responses which sometimes are frustrating.

**Background**

Intermountain Healthcare is a not-for-profit integrated healthcare delivery system based in Salt Lake City, Utah. It provides healthcare for the entire state of Utah and parts of southeastern Idaho. Intermountain maintains 22 inpatient hospitals (including a children’s hospital, an obstetrical facility and a dedicated orthopedic hospital), more than 185 outpatient clinics, and 18 community clinics serving uninsured and low-income patients. Intermountain provides primary and specialty care for approximately half of the residents of the state of Utah. Intermountain Healthcare has been recognized in the literature for supporting best care practice with clinical decision support interventions (13). Intermountain’s infobutton manager (a key component of this effort) is used regularly and its development and uptake have been detailed previously (14).

Infobuttons have been in use at Intermountain Healthcare for over 15 years (15,16). They have been integrated in two separate clinical systems, including usage from 4 major modules within these systems. Usage has steadily increased over the years, with over 1,700 unique monthly users, accounting for over 18,000 infobutton sessions per month.

Infobuttons have been made available for patient use from Intermountain’s patient portal MyHealth since early 2014. MyHealth is actively used by many patients, with tens of thousands of logins per month. This implementation based on the general OpenInfobutton (15) has been augmented with local services, supporting internal logging, integration with terminology services, and enhancements to satisfy local security requirements. This is the first exposure of infobuttons to the public network in Intermountain Healthcare. Infobuttons are available from the health concerns, medications, and lab value modules in myHealth. Currently, infobuttons link to seven different e-resources from MyHealth, including Krames StayWell, MedLine Plus, HealthFinder.gov, FamilyDoctor.org, Mayo Clinic, Drugs.com, and LabTestsOnline.

This research analyzes infobutton responses in MyHealth from three modules including Medications, Tests and procedures, and Health concerns. The analysis is based on 400 sampled concepts from each module. Specifically, we focus on automatically detecting empty responses for these concepts in infobutton responses and report the related statistics in pursue of enhancing the current implementation of infobuttons.

The targeted concepts are sampled from the Enterprise Data Warehouse (EDW) where patients’ health records are saved. In each module, we select the two sets of concepts, which are the top 200 and the bottom 200 ranked by the usage of all the concepts. The selection of these two sets enables us to differentiate infobutton performance between the most common conditions and the least common conditions. We expect that infobutton supports the common conditions better than the uncommon conditions.

The empty responses will be recorded and sorted out by module and by e-resource respectively. By Module, since Medication has over 75% of infobutton usage from the previous research, we expect that infobuttons will perform...
well in Medication with less empty responses than the other two modules. Also, by classifying empty responses in each e-resource, we could evaluate the performance of e-resources and expect variances in performances of e-resources.

In addition, the research compares empty responses in e-resources which supports HL7 standard with these without HL7 standard. We expect that the HL7 standard API results in more accurate search items as well as less empty responses because HL7 standard is context-aware with the patient information as well as the matching criteria are based on the standard coding systems, instead of free text.

Methods

This section describes the complete process of retrieving infobutton links and detecting empty responses from all the infobutton supported e-resources in three modules of patient portal MyHealth. The process includes building a tool, sampling concepts, and collecting empty responses from each e-resource. The tool automatically sends an infobutton request for each sampled concept and record empty responses if detected. The concepts are sampled from Intermountain’s enterprise data warehouse, which stores longitudinal patient data. The empty responses are collected by running the tool against the sampled concepts.

Automatically detect empty responses

We designed a program to analyze the content of e-resources derived from infobutton requests in MyHealth. In Figure 1, we demonstrate the workflow of the automatic process. First, a series of concepts is fed into the program for analysis. For each input concept, an infobutton request is composed, using the context information from a test patient and a test provider (which were both predefined and fixed for all the infobutton requests). The infobutton manager processes the request and generates a corresponding response with supported e-resources. For each URL listed from the supported e-resources, the program runs the URL and retrieves and parses the remote content. By parsing the header of the content we record the status code from web pages while the number of returned items from each URL is recorded by parsing the body of the content. A set of rules leads the process of parsing the remote content. In a rule, we define the path in HTML/XML to extract interested information from a specific e-resource. Each e-resource has its own set of rules that allow us to find and pinpoint key information in the page for analysis. By these predefined rules, our program is intelligent enough to identify empty responses and record URLs and other metadata for further analysis.
We extracted a large data set for analysis from each of the three infobutton domains. We created two sets of concepts for each infobutton supported modules including medications, health concerns, and tests and procedure. In a module, these two sets represent the most common conditions and the most uncommon conditions based on usage of concepts. In this research, we selected the top 200 to represent to most common conditions while the bottom 200 to represent the uncommon conditions. For each sampled concept, we compose an infobutton request. In Table 1, we display the count of hits for the concepts in the top and bottom set respectively in each module in year 2014. In Table 2, we demonstrate the representations of concepts from the top 5 and the bottom 5 in each module separately, where by (T) it refers to the top 200 concepts and by (B) it refers to the last 200 concepts ranked by the count of concepts’ hits in EHR system.

<table>
<thead>
<tr>
<th></th>
<th>Health concerns</th>
<th>Tests and procedures</th>
<th>Medications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>865,039</td>
<td>58,606,872</td>
<td>2,446,961</td>
</tr>
<tr>
<td>Bottom</td>
<td>200</td>
<td>527</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 1. The counts of top 200 and bottom 200 concepts in modules.
Table 2. The representations of the top 5 and bottom 5 concepts in modules.
(T): top 200 concepts; (B) bottom 200 concepts.

<table>
<thead>
<tr>
<th>Health concerns (T)</th>
<th>Health concerns (B)</th>
<th>Medications (T)</th>
<th>Medications(B)</th>
<th>Tests and procedures (T)</th>
<th>Tests and procedures (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal pain</td>
<td>Chronic obstructive pulmonary disease exacerbation</td>
<td>Norco</td>
<td>benzoyl peroxide</td>
<td>Complete Blood Count</td>
<td>Protein Electrophoresis</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>Addison's disease due to autoimmunity</td>
<td>Percocet</td>
<td>Serenagen</td>
<td>Chemistry 12 Panel</td>
<td>Adenosine Deaminase</td>
</tr>
<tr>
<td>Hypertension</td>
<td>Retinoblastoma</td>
<td>Fionase</td>
<td>Tyrex-1</td>
<td>CBC, Differential</td>
<td>Acid Ham's Test</td>
</tr>
<tr>
<td>Depression</td>
<td>Cystic Kidney Disease</td>
<td>Ibuprofen</td>
<td>sumatriptan</td>
<td>Chemistry 7</td>
<td>Serum or Plasma Cholesterol/HDL</td>
</tr>
<tr>
<td>Anxiety</td>
<td>Right to left shunt due to patent ductus arteriosus (PDA)</td>
<td>Augmentin</td>
<td>Levothyroxine</td>
<td>Lipid Profile</td>
<td>Beta Melanocyte Stimulating Hormone</td>
</tr>
</tbody>
</table>

Result from auto-detection of empty responses

We ran this tool for auto-detecting empty responses against the top and bottom 200 concepts in the module of Health concerns, medications, and Tests and procedures. Each concept sends an infobutton request. The process resulted in a report of empty responses for all the e-resources. As shown in Figure 1, the last step involves the capture of relevant output from the auto-detection of empty responses. These records includes information such as calling module, concept, the name of website, number of returned items from the website, status code of http response from the website, and the URL if the link does not return any items for the query. In our research, we processed and stored all 'empty responses' from the e-resources linked to from infobuttons in MyHealth, including Family Doctor, MedlinePlus, Krames, Health Finder, Mayo Clinic, Drugs.com, Lab Tests Online.

The number of empty responses was classified per e-resource and per module separately. In Table 3, we report the number of empty responses from each module as display in rows and from seven e-resources as shown in columns. In each row, a module contains the two sets of concepts represented by (T) for the top 200 concepts and (B) for the last 200 concepts ranked by the count of concepts’ hits in EHR system, as described in the Section of Dataset. The modules include health concerns, medications, and tests and procedures. The symbol ‘-’ in the table means that the e-resource in the column does not support the according module in the row from the initial design. In Figure 2, we visualize the result from Table 2, displaying the number of empty responses (vertical) for each e-resource in the top and bottom 200 concepts in three modules (horizontal). Health concerns has five e-resources, medications has four e-resources, and tests and procedures has 3 e-resources.

By module, the health concerns (B) has the most number of empty responses while the medication (T) has the least number. By e-resource, Drugs.com has 0 empty responses for all the infobutton requests while Health finder has the most of empty responses. We will further analyze the result in the next section.

Table 3. Total number of empty responses from each e-resource in three modules
(T): top 200 concepts; (B) bottom 200 concepts.
We also record the number of returned items from each e-resource. Table 4 shows the total number of returned search items from e-resources given the top and bottom 200 concepts in three modules. The result will be further analyzed in the next section.

<table>
<thead>
<tr>
<th></th>
<th>Family Doctor</th>
<th>Medline Plus</th>
<th>Krames</th>
<th>Health Finder</th>
<th>Mayo Clinic</th>
<th>Drugs.com</th>
<th>Lab Tests online</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health concerns (T)</td>
<td>6386</td>
<td>8289</td>
<td>4000</td>
<td>983</td>
<td>116968</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Health concerns (B)</td>
<td>513</td>
<td>12815</td>
<td>10598</td>
<td>139</td>
<td>19813</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Medications (T)</td>
<td>-</td>
<td>3202</td>
<td>7611</td>
<td>-</td>
<td>26012</td>
<td>3065705</td>
<td>-</td>
</tr>
<tr>
<td>Medications (B)</td>
<td>-</td>
<td>8722</td>
<td>11141</td>
<td>-</td>
<td>9527</td>
<td>11577212</td>
<td>-</td>
</tr>
<tr>
<td>Tests and procedures (T)</td>
<td>-</td>
<td>-</td>
<td>13215</td>
<td>-</td>
<td>15573</td>
<td>-</td>
<td>3804</td>
</tr>
<tr>
<td>Tests and procedures (B)</td>
<td>-</td>
<td>-</td>
<td>16806</td>
<td>-</td>
<td>5517</td>
<td>-</td>
<td>362</td>
</tr>
</tbody>
</table>

Table 4. Total number of returned items from each e-resource in three modules (T): top 200 concepts; (B) bottom 200 concepts.

Analysis of empty responses

The empty responses from websites were categorized by module and by e-resource. From the result, we evaluated the performance of infobutton by module, by e-resource, and by support for the HL7 standard (or the lack thereof).

An e-resource can be configured to support only one or multiple modules. For example, Krames supports all the three modules while Lab Tests Online only supports tests and procedures. In our results, per concept calculation accounts for only the supported concepts in supported modules for an e-resource in order to normalize for comparing performance of different e-resources that have different coverage.

In Table 5, we demonstrate the rate of empty responses and the average number of returned items from each e-resource. Each column refers to an e-resource where the last column reports the average over all the e-resources. In the row for rate of empty responses, for each e-resource, the rate is computed from all the infobutton requests in the process by:

\[
\text{rate of empty responses per eResource} = \frac{\text{total number of empty responses per eResource}}{\text{total number of infobutton visits to the eResource's website}}
\]

The average of the return items per concept for an e-resource is computed by:

\[
\text{average of returned items per request} = \frac{\text{total number of returned items per eResource}}{\text{total number of infobutton visits to the eResource's website}}
\]

The average rate of empty responses for all the e-resources is 30.60%. HealthFinder.gov had the highest rate, which almost doubles the average rate for all the e-resources. Drugs.com and Krames have the lowest overall ‘empty
response’ rate. Drugs.com has a zero empty responses because of its large database as well as possibly broader matching criteria, which can be inferred from its average number of the returned items, at 37,450 per requested concept. That average number is much bigger than the rest of the e-resources. By taking out Drugs.com from the average returned items, the average number of returned items for all the e-resources changes from 5,407 to 57 per requested concept. Krames has low rate at 1.00% of empty responses and a moderate number 54 for the average number of search items, which is close to the average number by counting out Drugs.com. Health Finder returned few search items but it has a high rate of empty responses, which indicates that the website might have less coverage of concepts. Mayo Clinic has a lower than average of the empty responses but has a bigger number of returned items, which means that it has a good coverage of content and therefore is more likely to return results.

<table>
<thead>
<tr>
<th>Rate of empty responses</th>
<th>Family Doctor</th>
<th>Medline Plus</th>
<th>Krames</th>
<th>Health Finder</th>
<th>Mayo Clinic</th>
<th>Drugs.com</th>
<th>Lab Tests online</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>51.75%</td>
<td>24.13%</td>
<td>1.00%</td>
<td>60.50%</td>
<td>27.08%</td>
<td>0.00%</td>
<td>49.75%</td>
<td>30.60%</td>
<td></td>
</tr>
<tr>
<td>Avg returned items</td>
<td>36</td>
<td>59</td>
<td>54</td>
<td>7</td>
<td>224</td>
<td>37,450</td>
<td>21</td>
<td>5,407</td>
</tr>
</tbody>
</table>

Table 5: Rate of empty responses and average returned items per e-resource

In Table 6, we report the percentage of empty responses for each e-resource by sets of concepts. The concepts are sampled from the top 200 and bottom 200 and grouped from all the supported modules. For example, Krames supports three modules so its top concepts cover the top 200 from all the three modules. For an e-resource, the percentage in the table is computed by the sum of empty responses over the total number of concepts in the top and bottom concepts separately. Noticeably, the top concepts, with an average rate of empty responses at 14.69%, always have lower percentage of empty responses comparing to the bottom concepts, with an average rate of empty responses at 46.51%. That means, all the e-resources support the most commonly used concepts better than the least used ones. Drugs.com has 0 empty responses while Krames and MedlinePlus have the lowest rate of empty responses. But MedlinePlus has a significant number of empty responses for the bottom concepts, comparing the Krames. Health Finder and Family Doctor have both the highest rate of empty responses in the bottom concepts, where the rate is 85.50% and 86.00% respectively. With that high rate of empty responses, we could suggest removing the Health Finder and Family Doctor from the list for these bottom concepts and have users focus on these which return non-empty items.

<table>
<thead>
<tr>
<th>Top concepts</th>
<th>Family Doctor</th>
<th>Medline Plus</th>
<th>Krames</th>
<th>Health Finder</th>
<th>Mayo Clinic</th>
<th>Drugs.com</th>
<th>Lab Tests online</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.00%</td>
<td>6.50%</td>
<td>0.33%</td>
<td>35.00%</td>
<td>11.00%</td>
<td>0.00%</td>
<td>32.00%</td>
<td>14.69%</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Empty responses from e-resources in the top and bottom 200 concepts

In Table 7, we present the rate of empty responses for each e-resource by module. The symbol ‘-’ means that the e-resource in the column is not supported in the module in the row. For an e-resource, the rate of empty responses is computed by the number of concepts with empty responses over the total number of concepts in a module. The medication has the lowest rate of empty responses at an average of 10.42%, comparing to the highest of 35.20% in Health concerns. Medication has the most usage from previous research of infobutton usage. That suggests that the e-resources provide better coverage in modules with higher infobutton usage. Krames is supported in all the three modules and performs well, where Krames has the lowest rate of empty responses in all of the modules. Health Finder and Family Doctor both support a single module of Health concerns but have a high rate of empty responses, where is at 60.50% and 51.75% respectively. We would suggest lower the accessing priority of these two e-resources in Health concerns so people will have less chance to be frustrated with empty responses. Medline Plus and Mayo Clinic perform similarly in Health concerns and Medications, which are both close to the average of all the e-resource in each module. Lab Tests online has a high rate of empty responses, so we would suggest to lower the accessing priority for Lab Tests online while the vender could try to broaden the search criteria as well as improving content coverage in the backend.
The difference in rate of empty responses between e-resources with and without a HL7 standardized API is sharp, as shown in Table 8. The rate of the empty responses is computed by the number of empty responses grouped by the e-resource’s API over the total number of infobutton requests fired from that group. In Table 8, we compare the rate of empty responses between e-resources with HL7 standard API and non-HL7 standard API, in each module with the top and bottom concepts. All the e-resources have a very low rate of empty responses for the top concepts in both HL7 standard and non-HL7 standard column. However, the e-resource supporting HL7 standard API has a noticeably different performance in Health concerns and Tests and procedures. In Medication module, the performance of infobutton requests is generally the best for both the HL7 standard API and the non-HL7 standard API, where the rate of vacancy is similarly low. The non-HL7 standard API, especially Drugs.com which has zero empty responses, is mostly based on free text search and has adjustable broader matching criteria, which results in a lower rate of empty responses. Overall, the bottom 200 concepts using the non-HL7 standard API in the Health concern have the highest rate of empty responses. Interestingly the top 200 concepts also using the non-HL7 standard have the lowest rate of empty responses, mostly because of Drugs.com’s zero contributions to the empty responses in this module. Despite the contribution of the non-HL7 standard API from Drugs.com, the HL7 standard API performs similarly to the rate of empty responses in Medication.

<table>
<thead>
<tr>
<th>Module</th>
<th>HL7 standard API</th>
<th>Non- HL7 standard API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health concerns (T)</td>
<td>5.50%</td>
<td>20.17%</td>
</tr>
<tr>
<td>Health concerns (B)</td>
<td>28.50%</td>
<td>74.50%</td>
</tr>
<tr>
<td>Medications (T)</td>
<td>1.00%</td>
<td>0.75%</td>
</tr>
<tr>
<td>Medications (B)</td>
<td>15.00%</td>
<td>14.50%</td>
</tr>
<tr>
<td>Tests and procedures (T)</td>
<td>1.00%</td>
<td>28.00%</td>
</tr>
<tr>
<td>Tests and procedures (B)</td>
<td>1.50%</td>
<td>58.00%</td>
</tr>
</tbody>
</table>

Table 8: Rate of empty responses in e-resources supporting HL7 API and non-HL7 API

**Discussion**

This paper analyzes the empty responses from infobutton manager by module, by e-resources, and by supporting HL7 standard or not.

The common concepts always have less empty responses than the uncommon concepts. All the e-resources perform better at returning content for the common concepts than the rare ones. We conclude that the overall content coverage for common concepts is good. In the meantime, more effort should be sent to improving the content coverage of the less commonly used concepts.

By module, medications has the best performance in terms of returning less empty responses and more number of returned items for a requested concept. That matches the fact that medications has the highest usage of infobuttons. As a module with higher demand, we conclude that infobuttons do well support the usage.

By e-resource, Krames is notably the best e-resource among the seven e-resources we analyze in this paper. Krames has the lowest rate of empty responses and returns a moderate number of items. HealthFinder and Family doctor will need to improve significantly upon the number of empty responses in order to be viable e-resources on the infobutton. Drugs.com never brings users to a blank page, but leaves some questions about users’ perceptions with the content returned. Further analysis will be needed to assess the quality of the returned items.
E-resources supporting HL7 standard consistently perform better than the non-HL7 standard e-resource by a significant margin. As such, we would encourage implementers toward the usage of the HL7 standard for e-resource providers as possible.

Conclusion and Future work

We present a method to analyze the content from infobutton requests. The empty responses are automatically detected from e-resources. The result of the empty responses is analyzed by module and by e-resource. We suggest that more effort will be needed to cover content from the less common conditions, such as less used concepts and less used modules. We also encourage e-resource providers to implement the HL7 standard therefore the performance of infobuttons can be improved.

In future, we will need to analyze the quality of the content from different e-resources. Also we will need to read through the empty responses and categorize the reasons for the failures. Furthermore, we can build a tool to analyze the content offline and dynamically create the list of e-resources promoting these with better content and therefore avoid the empty responses at real time.

References

A Novel Multiple Choice Question Generation Strategy: Alternative Uses for Controlled Vocabulary Thesauri in Biomedical-Sciences Education

Marcelo A. Lopetegui1,2, MD,MS, Barbara A. Lara1, MD, MPH, Po-Yin Yen1, RN, PhD, Ümit V. Çatalyürek1, PhD, Philip R.O. Payne1, PhD

1Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA  
2Departamento de Informática Biomédica, Clínica Alemana de Santiago, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, Chile

Abstract

Multiple choice questions play an important role in training and evaluating biomedical science students. However, the resource intensive nature of question generation limits their open availability, reducing their contribution to evaluation purposes mainly. Although applied-knowledge questions require a complex formulation process, the creation of concrete-knowledge questions (i.e., definitions, associations) could be assisted by the use of informatics methods. We envisioned a novel and simple algorithm that exploits validated knowledge repositories and generates concrete-knowledge questions by leveraging concepts’ relationships. In this manuscript we present the development and validation of a prototype which successfully produced meaningful concrete-knowledge questions, opening new applications for existing knowledge repositories, potentially benefiting students of all biomedical sciences disciplines.

Introduction

Multiple choice questions (MCQ) have been widely used as an assessment tool in biomedical sciences education, currently being a major component of recognized tests such as the Graduate Record Examinations (GRE)1, the United States Medical Licensing Examination (USMLE)2, and the American Board of Medical Specialties3. Appropriately constructed, MCQ can be used to objectively assess all levels of learning of the Bloom’s taxonomy of cognitive learning4, from concrete-knowledge up to application5.

A MCQ usually consists of a question or a statement to solved (usually referred to as “stem”), followed by a list of possible answers to choose from. There is only one correct answer, while the incorrect answers are distractors. Formulating MCQ, specially finding meaningful distractors, is resource intensive and time consuming. This, prevents educators from openly sharing question-banks with students, usually restricting access to local cohorts6. This behavior limits the contribution of MCQ to evaluation purposes mainly, while it is recognized that they can contribute to the student training process by improving knowledge retention and learning.7 Furthermore, it has been reported that exposing students to questions before receiving learning material can have beneficial effects on students’ learning8.

Thus, in an attempt to facilitate MCQ generation and expand their use beyond evaluation purposes, automatic MCQ generation approaches has been attempted.9 Although applied-knowledge questions require a complex formulation process10 (e.g., “what is the best antibiotic for a patient with infection X and allergic to Y?”), automatic MCQ formulation strategies have been able to successfully generate concrete-knowledge questions. Distinct approaches range from entirely automatic generation of true and false statements by leveraging classes and relationships from domain ontologies, and presenting those statements as possible answers for the stem “choose the correct sentence”9, up to complex computer-aided natural language processing (NLP) based methods11.

Expanding the previous efforts in automatic MCQ generation from domain ontologies, we envisioned a novel approach that automatically leverages existing knowledge repositories, aiming at providing open MCQ banks for students to train and learn. We hypothesize that a simple generation strategy based on concepts’ definition and their hierarchical relationships can be leveraged to generate MCQ with meaningful distractors. In the current manuscript, we describe the development and validation of our novel approach based in a controlled vocabulary thesauri, and present a working prototype: a new resource for students in biomedical sciences to learn and review concepts.
Methods

Central paradigm. We attempted to create concrete-knowledge questions, specifically definitions, relying on the premise that hierarchical relationships of concepts could be leveraged to retrieve meaningful distractors. Based on a hierarchical tree structure of concepts and definitions (i.e., a taxonomy), we can select a concept, and expose its definition as the “question”. Then, the concept name becomes the correct answer. Since concepts that share a parent concept are similar but at the same time mutually exclusive, they can be considered appropriate distractors. Thus, we then search for siblings in the hierarchy to retrieve distractors and complete the question (see Figure 1).

Data source. Ontologies, particularly in the biomedical domain, are well maintained repositories of concepts and relationships, suitable for our aim. We began our search for an appropriate source for questions’ generation in the Unified Medical Language System (UMLS) definitions table (MRDEF.RRF), which encompasses over one million definitions contributed by the National Cancer Institute metathesaurus (50.1%), the Gene Ontology (27.4%), the National Library of Medicine’s Medical Subject Headings (21.2%), the Foundational Model of Anatomy ontology (0.9%), and the Systematized Nomenclature of Medicine - Clinical Terms (0.5%). Out of the three major contributors, we selected the National Library of Medicine’s controlled vocabulary thesaurus (MeSH) due to its broad scope within the biomedical sciences, its rich definitions, the general good acceptance and recognition among the scientific community, and because of its interesting category-based hierarchical tree schema, which would allow us to easily navigate distinct disciplines.

Data representation. After accepting a memorandum of understanding, MeSH provides universal free access to the descriptors (or subject headings, hence the name) in both XML and ASCII format, and to the tree structure in ASCII format. The MeSH 2014 tree has 55,611 nodes, representing 27,983 unique descriptors from 16 categories (any given descriptor may be represented more than once in the hierarchical tree). For example, the descriptor “Respiratory Tract Fistula” is represented twice in the hierarchy: in location C08.702 and location C23.300.575.687. “C” stands for “Diseases”, “C08” for “Respiratory Tract Diseases” and “C08.702” for “Respiratory Tract Fistula”. Same wise, “C23.300” stands for “Anatomical Pathological Conditions”, “C23.300.575” for “Fistula” and “C23.300.575.687” also for “Respiratory Tract Fistula”. Most of MeSH descriptors contain a short free-text narrative, the scope note, giving the scope and meaning of the concept written by the MeSH team, sometimes referencing specific sources.

We transformed the path enumeration format of the MeSH tree and created an adjacency list model and a nodes table. Then, from the concepts table, we extracted each MeSH Heading (the descriptor name), the associated MeSH Scope note and the MeSH tree numbers (the location or locations within the hierarchy), and loaded everything onto a relational database (MySQL).

Algorithm. Out of the 27,983 concepts contained in MeSH, only 26,144(93.4%) had a definition (the “scope note”), reducing the number of useful initial nodes to 54,148. We did not delete concepts without definitions since even if they are not suitable for question generation, they can be used as distractors. We designed the algorithm following the simple approach stated above (Figure 1): it selects a random node from the tree and
exposes the definition of the node as the question. One of the choices is the name of the node being displayed (correct answer), while the other alternatives (distractors) are retrieved by looking for siblings within the sub tree.

Education-related research has shown that 3 options MCQ (2 distractors) provide a similar quality of a test as compared to those with 4 or 5 options\(^{12,13}\) which could improve efficiency in question generation. For our prototype, time was not a concern, but the available number of siblings was important. Table 1 shows the number of nodes with x number of siblings. For example, 36,321 nodes (67.1% of the tree) have at least 3 siblings, and thus could produce a question with 4 alternatives. Since the recommendations of using 3 alternatives MCQ comes from human generated questions, we thought to increase it to 4 to reduce the chance of selecting the correct answer by guessing while still being able to use 67.1% of the tree.

When more than 3 siblings are available, distractors retrieval occurs at random and the order of the choices (distractors + correct node name) are alphabetically sorted before being presented to the user.

### Application development.
In our preceding study concerning medical education (MoCK Test, manuscript under preparation), we parsed an existing open question bank, only available as a flat web page, and developed a native mobile application providing mobile-optimized access for students to test their knowledge on the go. We described the wide adoption and usage patterns, evidencing the benefits of mobile-optimized content: it allowed users to study whenever they had a short opening in their busy and interrupted life.

Based on those results, and our belief that learning tools should be available for all students regardless of their device preference or operating system, we named the new tool “MoCKTest 2.0” and developed it as a web based application, following the mobile first paradigm. We used free web technologies, including the Foundation (ZURB)\(^{14}\) framework on the client side, and PHP and MySQL on the server side. The responsive web design we used made our concept available to any internet capable device, while ensuring a mobile optimized presentation of the content.

### Evaluation and user acceptance.
Because biomedical science students were thought as the group that would most likely benefit from the content of MoCKTest 2.0, we invited medical, nursing and pharmacy students to try the application and participate in the evaluation section of our study. We contacted Ohio State University students from those disciplines via their institutional weekly newsletter and/or via personal referral, while access to the app was open to anyone willing to participate in the study (previous acceptance of the informed consent). A likert-scale based usability survey was created to evaluate user acceptance, consisting of 15 questions measuring concepts including ease of use, usefulness, satisfaction, and intention to use. The survey was triggered after users completed 15 questions, and participation was voluntary. In addition, we conducted semi-structured interviews with subject matter experts (SME) in biomedical education to examine and comment on the quality and meaningfulness of questions generated by MoCKTest 2.0, the potential benefits that this tool could provide to students as well as suggested improvements. The Office of Responsible Research Practices (ORRP) at The Ohio State University determined that this research protocol was exempt from IRB review, as it corresponds to a review exemption category established by federal regulations. This protocol was approved as such by the ORRP.

### Results

#### Prototype.
Our algorithm successfully generated questions using the proposed approach. The adjacency list model used to represent the tree resulted in good performance for question generation. We designed an appealing and intuitive user interface, and ensured mobile-optimized content with the responsive design provided by Foundation. The final prototype, MoCKTest 2.0, was successfully deployed in our production environment at the OSU Wexner Medical Center IT servers, and can be accessed at [http://www.mocktest2.com](http://www.mocktest2.com).
In an effort to minimize barriers to adoption, we implemented social login capabilities using OpenAuth standards[15]. This approach eliminates the requirement to create a user and remember a new password. It relies on permissions granted by the users to get their email address from their preferred social account, which is then used for authentication purposes. Users can choose to login with Facebook, Google or Microsoft accounts.

On application launch, users can select one of the MeSH categories and subcategories and start answering questions concerning the selected topic (see Figure 2). Due to the considerable size of the question bank (36,000+ concepts) and the randomness of question generation, we provide a favorite feature, allowing users to tag questions to practice later. By default, the app also permits students to answer questions they have gotten wrong in previous attempts. After each answer, the user receives immediate feedback, informing them which choice was the correct one, in case they selected a wrong answer (Figure 3). New question requests happen on the background via an AJAX call without page refresh, improving user experience and decreasing traffic to/from the server. Users can also assess their performance and get an overview of the number of questions answered per sub-category and the percentage of success on each of them. Stats can be reset by students to begin a new study cycle at any point.

**Users’ perception of the tool.** Invitations to participate in our study were sent to the students of The Ohio State University through their college specific newsletter of the second week of February 2014. In three weeks, 325 unique visitors (not exclusively Ohio State University students) viewed the landing page, while only 120 accepted the informed consent and created an account.
(conversion rate of 36.9%). Of those who used the tool, 50 answered the survey (41.7% response rate). Seventy eight percent found it useful for biomedical knowledge self-assessment, while 75% agreed that using the app could improve their biomedical knowledge and be useful for their medical, nursing or pharmacy education. Sixty seven percent believed that using the system could improve their performance in school. Ninety seven percent of respondents believed that learning to navigate the app was easy. A complete report of the survey results can be found in Appendix 1.

Feedback from educators. We interviewed four subject matter experts (SME) in biomedical education: a Vice Dean for Education and Associate Vice President for Health Sciences Education of a College of Medicine, an Assistant Dean for Prelicensure Programs and Professor of Clinical Nursing of a College of Nursing, and an instructor of licensure review courses including the National Council Licensure Examination for Registered Nurses (NCLEX-RN) and the National Council Licensure Examination for Practical Nurses (NCLEX-PN). During 15 minutes of detailed assessment of random questions, SME evaluated the question generation process and the quality and meaningfulness of the questions presented, based on standard MCQ item-writing guidelines16. Despite an overall satisfaction with the quality of distractors generated, two scenarios raised concerns with the stems. The first scenario corresponded to high-level nodes (close to the root), which seemed to produce vain and basic questions (i.e. “Tumors or cancer of the uvea.”. Answer: “Uveal neoplasms”). The second scenario corresponded to questions where, despite an appropriate level of complexity, the user was able to “guess” the correct answer. This seemed to happen when a variation of the node name was present in the definition itself (i.e. synonyms). The application received a consensual very good feedback on ease of use, and most importantly, they all agreed on the benefit of providing such a question bank to students for training purposes.

Discussion

Although adding complex post-processing for better articulation of the questions might improve the prototype, the novelty of our proposal relies on the simplicity of the algorithm presented. By relying only on a hierarchical structure and definitions, our approach is discipline agnostic and allows the use of any knowledge representations meeting these requirements.

Due to the nature of the datasets involved and the scope of the questions, our tool might not be used by teachers to generate questions for tests. However, the easy implementation of our approach and the open nature of the content, provides an unprecedented platform for students to train, learn and self-assess their knowledge. Despite the two conflicting scenarios identified, the tool as is, presents itself as a contribution to education for all biomedical students.

Overcoming conflictive scenarios. Regarding the first scenario where “high level nodes” might be too general and produce vain questions, a potential solution corresponds to limiting questions retrieval to certain levels of the tree. However, each MeSH category has different number of branches (sub categories), and each “branch” has different levels of depth [Figure 4]. For example, the subcategories of “Publication Characteristics” have only five levels of depth, while one of the subcategories of “Organisms” reaches twelve levels of depth. For that reason we believe that level 4 of “Publication Characteristics” might not be comparable in complexity to a level 4 question from Organisms. We hypothesize that a meaningful way of overcoming this issue might be to retrieve nodes from up to a certain level away from the leaves (bottom up). For example, a level 8 question from the geography category (deepest level) is as granular as that category can go, and thus might be equivalent in complexity to a level 11 question of Chemical and Drugs category (deepest level). A complementary approach would be to trim out categories that are expected to produce futile questions, such as “publication characteristics”. Future work will focus on validating these hypotheses.

Although the second scenario - where the user could guess the correct answer- might be regarded as a limitation of the tool for evaluation purposes, we believe that this event could also be seen as a learning opportunity. The end goal of our tool is to empower the user to learn, not to serve as an evaluation tool. Thus, even if the student is able to guess the answer to a question he would have missed, he is actually learning the concept by reading the definition and associating it to the correct answer. Moreover, many times the definitions
contained in MeSH provide more detailed information than what would be strictly necessary to answer the question, thus providing more content to this learning opportunity.

**Ongoing efforts.** Based on feedback received during the interviews, we envisioned possible improvements for the prototype. First, we will provide the ability to read the definition of distractors when receiving feedback on any given question, thus increasing the learning opportunities for unknown domains. Second, we will provide the user with module for navigating down the tree and discovering definitions of unknown concepts, akin to the MeSH browser\(^\text{11}\). Another interesting idea proposed by the nursing experts corresponds to an alternate question generation process, also based on the hierarchical structure available. The model proposes to create “except” questions by selecting a random parent node, presenting the name of the node as the group name, and listing the siblings plus a random node retrieved from a distant relative. Thus, the user will be asked to identify the term that doesn’t belong to the group. For example: All of the following correspond to Peroxisomal Disorders EXCEPT: Adrenoleukodystrophy, Mevalonate Kinase Deficiency, Refsum Disease, Fanconi Syndrome and Zellweger Syndrome. Although the idea seems interesting, it requires further testing and tuning, since the meaningfulness of these approach might be highly related and affected by the depth of the nodes.

**Limitations.** Although we successfully validated the question generation strategy, the use-case as an educational tool for biomedical students might be limited due to the data source selected: MeSH might not be the best source of comprehensive medical knowledge, and may have several biases in the concepts included/excluded in this hierarchy. Our evaluation of the prototype seems to suggest that the strategy might be useful for medical education, although a more large-scale evaluation effort is probably needed. Future efforts will focus on expand the tool to a larger population to gain feedback. The persistent data storage implemented in our prototype is not

---

**Figure 4:** Composition of MeSH. The central pie chart represents the relative contribution of each category to the tree. The orbital graphs represent each category composition (distinct number of subcategories, and distinct levels of depth). For example, “Geographical” has only one subcategory, with 5 levels of depth, while “Organisms” has 5 subcategories, varying from 4 levels down to 11 levels of depth). The size of a node represents the relative number of concepts in that level.
the most efficient in this context. A graph database would improve the queries to lookup distractors. The approach used might become limited if the project scales to large populations.

**Conclusion**

We introduce MoCKTest 2.0: a new asset for students in biomedical sciences to learn and review concepts. Our simple tool leverages existing resources such as controlled vocabulary thesauri, creating concrete-knowledge questions and opening a new educational use for knowledge repositories. Students and educators recognized the contribution of such a question bank to the learning process, extending multiple choice questions contribution beyond evaluation purposes. Despite minor limitations, our idea has the potential to contribute to the training and education of scientists and researchers of all the biomedical sciences.

**References**


Appendix 1: Diverging stacked bar charts of survey results. This visualization easily allows to identify the skew between total positive and negative responses, due to the central base. Each question is represented as a row. The total width of the bar shows the percentage of respondents who have non-neutral feelings towards the statement. The depth of color represents the intensity of feeling.

### Ease of use

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>The app is simple to navigate.</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Learning to use the app is easy for me.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>The app is easy to use.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>I find it difficult to work with the app.</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>I think that I could always remember how to use the app.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

### Usefulness

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I find the app useful for biomedical knowledge self-assessment.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Using the app makes it easier to acquire biomedical knowledge.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Using the app could improve my biomedical knowledge.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Using the app allows me to learn biomedical knowledge anywhere.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

### Satisfaction

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using the system could improve my performance in school.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>I find that the benefits of using the system are bigger than the effort of using it.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>I think the app can be useful for my medical/nursing/pharmacy education.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>I think the app can contribute to my clinical training.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

### Intention to use

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would like to use this system regularly.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>I would recommend others to use this app.</td>
<td>-10</td>
<td>-10</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>
Veterans Health Administration Experience with Data Quality Surveillance of Continuity of Care Documents: Interoperability Challenges for eHealth Exchange Participants

Jay Lyle, PhD, Omar Bouhaddou, PhD, Nathan Botts, PhD, Marie Swall, Eric Pan, MD, Terry Cullen, MD, Margaret Donahue, MD, Nelson Hsing, ScD*

US Department of Veterans Affairs, Washington, DC

Abstract

As part of ongoing data quality efforts authors monitored health information retrieved through the United States Department of Veterans Affairs’ (VA) Virtual Lifetime Electronic Record (VLER) Health operation. Health data exchanged through the eHealth Exchange (managed by Healtheway, Inc.) between VA and external care providers was evaluated in order to test methods of data quality surveillance and to identify key quality concerns. Testing evaluated transition of care data from 20 VLER Health partners. Findings indicated operational monitoring discovers issues not addressed during onboarding testing, that many issues result from specification ambiguity, and that many issues require human review. We make recommendations to address these issues, specifically to embed automated testing tools within information exchange transactions and to continuously monitor and improve data quality, which will facilitate adoption and use.

*Correspondence to: Nelson Hsing, Department of Veterans Affairs, Office of Interoperability (Nelson.Hsing@va.gov)

Introduction

The United States Veterans Health Administration (VHA) provides healthcare for almost nine million veterans. Increasingly, this care is provided in coordination with private sector providers. When care responsibilities are distributed among different providers, it is critical that all participants stay informed of each other’s activities and decisions so that care continuity can be maintained for the patient. Reliable quality data is needed for referrals and transitions of care: this is a key driver development of the VLER Health program. The Healtheway initiative, a public-private collaborative that promotes standardized and trusted exchange of health data among participating member organizations, provides the means through which the VA could effectively test this type of national-level health information exchange.

In 2012, the VHA participated in a pilot program to test the use of Healtheway’s eHealth Exchange (at the time known as the Nationwide Health Information Exchange, or NwHIN). This pilot program included VA, Department of Defense, and 12 extant Health Information Exchanges (HIEs) as part of the VA’s VLER program. Subsequently, in 2013 the project became operational, and now includes upwards of 45 exchange partners with whom over 80,000 Continuity of Care Documents (CCDs) have been shared. VA is currently exchanging CCDs based on the HITSP C322 and C62 formats, and has initiated the process of enhancing the infrastructure to incrementally migrate partners to the Consolidated CDA document specification (C-CDA).

Research previously conducted by the VLER team, based on the VLER Health pilot period, identified several important data quality findings, including issues with C32 conformance to the specifications, optionality in the specification that adds to interoperability problems, and evidence that current testing tools cannot effectively validate clinical data during clinical operation. D’Amore, et al. showed that data quality issues persist even after migration to the newer C-CDA standard and after EHR systems have been certified based on the Meaningful Use criteria. At the same time, we showed that even pilot phase data monitoring provided critical enhancements to the data quality. These findings suggested that ongoing data quality monitoring and quality surveillance would continue to enhance adoption, decision-making and the quality of care for veterans using multiple providers by continuing to identify potential areas of miscommunication and interoperability challenges.
Comparison of VA data quality evaluation results to those gathered by other researchers in this domain has identified similar classes of issues independent of the specific study settings (3). In this paper, with over a year of experience in production, we report specific data quality issues presented by C32 documents provided to the VA. These have informed a data quality surveillance framework that includes the use of automated tools and systematic scoring of received documents.

**Objectives**

We posit that there are two key goals for improvement of the current quality of exchanged data: 1) reliable automated data integration that supports semantic interoperability and 2) increased clarity and usability of the data when accessed by care staff. In the long term, health systems should support automated semantic interoperability, including rule-based alerts and automated integration of received data into the health record. It is widely recognized that interoperability and decision support standards are not yet mature enough to support large-scale implementation of this sort of functionality (incremental improvement to support such a goal being a motivator for this research). However, while we are not yet able to completely rely on automated interfaces for the management of clinical data, current health information exchange standards increasingly support consistent representation of health record data to human reviewers. Although the current operation does not automatically incorporate received data in the patient health record, it does present externally sourced documents to clinicians on demand. The near-term requirement, then, is that the data be comprehensible to the viewing clinicians. It is important to note that clinicians may review the data not only one source document at a time, but more naturally, as combined data aggregated from multiple source documents. The first option can be satisfied by the narrative block of the CCD, but the second option requires parsing of the structured entries, and thus requires that these structured entries share design assumptions to a degree approaching that required for automated integration.

Whether a document meets one of these two requirements for a given element does not imply that it meets the other: a document may be structured correctly but provide information that is confusing to read, or it may provide clearly comprehensible information in a file that fails to conform to the rules defined by the standard specification. The focus of this research is to identify cases where information may be confusing or incomplete to a human viewer so that the specification may be refined—ideally to the point where it may support even more automated functionality—and to identify opportunities to leverage automated standards conformance tools to assist in that effort.

**Methods**

We have identified several cases where incoming documents do not conform to the published specification. We provide these cases to the institutions from which they were received as feedback to support their document generation improvement efforts. We score these documents according to their richness (how many of the sections supported by C32 are sent for a given sample of patient cases), quality (percentage of present fields filled in correctly), and semantic interoperability (percentage of fields of coded data type that are provided in the specified standard terminology). We are in the process of automating this process, as it is extremely time-consuming to perform manually. The automation will also allow us to process a larger sample of documents, perhaps even every document received, and achieve more accurate scores that account for data quality variability from one patient record to another.

Of equal interest are documents where a field instance may conform to a specification but fail to provide unambiguous information. We also find that assessing these issues with an eye to automation facilitates unambiguous definition of quality criteria even for manual review.

We designed a two-stage process to address both questions. In the first stage, a document is tested for conformance to specification by automated tooling; in the second, the document is inspected for clarity in a human interface. This process identifies a majority of technical errors, and it also identifies clarity issues not detected by automated tools. Further, by conducting the automated tests first, we would pre-populate the manual scoring template, reducing the effort required and allowing the analyst to focus on otherwise undetectable semantic issues.

Figure 1 shows the process at a high level.
Referring to the numbering in Figure 1, incoming documents (1) are automatically scored by existing tools, such as Model Driven Health Tools (MDHT) (2), or by other products, such as the SMART CCDA evaluation tool (3). The results of these automated activities are output in a common format (4) and stored in a database (5). The database supports an analyst screen (6) for the document, in which an analyst can identify problems found while inspecting the document in the clinician’s VistaWeb interface (i.e., VA clinician user interface) (7); this input is also stored in the database (5). The resulting detailed database can then support any number of reports, scoring algorithms, or other analytical processes (8). These processes can even employ different standards of conformance; e.g., whether and when a particular conditional field should be considered required.

We encountered a number of issues in this effort, including most notably a shortfall in tooling functionality and difficulty in distinguishing the causes of clarity issues.

XML parsers can provide validation to the CDA schema. For example, the National Institute of Standards and Technology (NIST) CDA validation tools perform much of the validation to the template constraints; however, it
provides only limited support for assessing conformance to terminology constraints. And, while the validation that it provides is useful guidance for developers, it is not designed for monitoring quality in a production environment by detecting or counting issues over time, across multiple patient cases. We have defined requirements for the enhancement of the MDHT utility, but their implementation has not been completed.

Our process for assessing clarity relies on the transformation of the CDA xml document into a legible web page. This is performed by a transform—an XSLT template—that renders the xml as html for a web browser to display. The selection of this transform is significant. The CDA Release2 specification includes a simple transform, based on rendering the narrative blocks of the CDA-based document. The VA has built another type of style sheet that renders the information by parsing the structured entries. This more functional approach addresses several issues of optionality and alternate permissible patterns, as observed among VA exchange partners. Furthermore, as mentioned above, this method allows for aggregated views of data domains across multiple sources of data: e.g., one allergy list that consolidates data from the VA EHR and from external sources. These solutions are locally useful, but the CDA specification stipulates that a document cannot rely on a particular transform for faithful presentation. All enhancements built into the transform will have to be fed back to the standards development process (i.e., at HL7) to ensure a common understanding of these ambiguous points.

**Sampling**

We sampled documents retrieved by the VA from Exchange partners via the eHealth Exchange program in response to queries made by VA clinicians for the purpose of treatment. We tested from one to three documents from each partner, intending to support ongoing, operational monitoring of the quality of data in the eHealth Exchange. We selected documents that had already been requested and received: we were assessing actual operations, as experienced by clinician users, not requesting de novo data to test functionality. All analysis was conducted by trained analysts on systems behind the VA firewall to protect patient confidentiality.

Because we were interested in how participants populated their documents, we did not select files at random, but rather selected those document instances with the most data. As a result, the assessments do not represent an absolute proportion of well-formed data instances provided by a partner, but the proportion of defined elements for which the partner demonstrates the ability to form properly. If a partner had many documents with few sections populated and a few documents with many sections populated, we assessed the latter.

A similar interest guided our scoring within documents. We did not want a preponderance of data in a field with many instances to drown out the quality measure of a field with few instances.

**Error Classification**

Data quality errors were detected in three ways. First, required fields are counted as errors if missing; optional and conditional fields are not. Second, some errors can be detected by conformance tools, e.g., terminology conformance. And third, human review will identify cases where the document may conform to the technical specification but still fails to provide clear meaning to a reader.

We also endeavored to categorize errors by cause. The general assumption is that an error is the responsibility of the sender, and this is true for all conformance errors. However, there are three possible sources of clarity issues: 1) the author or sender, 2) the stylesheet used to present the data to the reviewer, and 3) in some cases the specification is not sufficiently detailed to be unambiguous, and consequently partners may make different assumptions, each of which conforms to the specification. For any clarity issue, we had to inspect the underlying xml code to determine which of these the case was.

If the issue seemed to be ambiguity in the specification, we identify the ambiguity so that stakeholders can propose new rules—whether to refine the standard specification or to share among eHealth Exchange participants who can then implement them in their own data quality monitoring.

Source errors are reported to the document sender. Transform errors are reported to the VA EHR GUI (VistAWeb) development team so that the assumption can be corrected. Ambiguities in the specification (and recommendations
for their resolution, if appropriate) are reported to the HL7 Structured Documents working group and to the Healtheway Specifications Factory workgroup.

**Scoring**

The partner richness score is a percentage defined as the average number of sections provided in the sample set of selected documents divided by the total number of sections defined by the specification.

The document quality score is the average of quality scores for all of the elements in the document. In a given document instance, each data element gets a score ($S_e$) consisting of the total error-free instances divided by the total expected instances.

$$S_e = \frac{\text{count of present & correct instances}}{\text{count of expected instances}}$$

For singular elements, this value is either one or zero, but for repeatable elements it is calculated. A document listing three allergies, but only two with reaction types (a required field for a C32 allergy), would receive a score of 0.67 for the reaction type data element. This way, we can capture effects of intermittent errors without ignoring evidence of appropriate integration.

The document quality score ($S_d$) is the average of these data element scores.

$$S_d = \text{Average} (S_e)$$

This average weights all elements equally, so that a heavily populated section does not have undue influence on our understanding of the sender’s capabilities.

The construction of the score in this compositional manner allows scores to be calculated not only for individual documents, but for any slice of information—partners, partner classifications (by size, geography, onboarding stage, etc.), data elements, sections, time periods, etc. It can be used not only for static scores, but also for trends, outliers, fraud detection, or any other perspective that might assist with the continuous improvement process.

The semantic interoperability score is the proportion of coded elements correctly formed, by element, not instance. This distinction is important because a document may contain a code that is incidentally correct. One document, for instance, uses “mg” for milligram, which appears to be the correct Unified Code of Unit of Measure (UCUM) code, but it also uses “K” for thousand, whereas the UCUM meaning of “K” is degrees Kelvin. This document is using local unit codes, some of which happen to be identical to the standard codes. If any errors are detected, all codes for the field must be considered suspect.

Because the tooling for evaluation is still under development, we only have preliminary quantitative assessment of the semantic interoperability score, based on manual counts. All counts are noted in a specially created Excel spreadsheet.

**Results**

![Data Quality and Interoperability Levels](image)
Figure 2: Example CDA Document Assessment Scores

Figure 2 shows an example partner assessment. The assessed partner’s documents included slightly under half the number of sections specified by the C32 (data richness = 45%). While overall quality scored 70%, coded elements as a group fared significantly worse (26%). Parenthetical figures indicate change from the partner’s previous scores.

Overall VLER partners richness scores ranged from 12% to 82%, with an average of 55% and median of 59%. When richness is low, it is unclear whether the sender is incapable of sending the missing sections or the data is simply not applicable to the patient in question. Automated operational inspection of greater volumes of documents should provide benchmarks for this judgment.

Figure 3: Histogram of CDA documents richness score across VA Exchange partners

The manual quality assessment process requires appreciable amounts of time. Due to the small sample size, we are able to provide only qualitative assessments of document quality. In addition, we have been unable to demonstrate a clear trend of quality improvement in incoming documents, given the short time range. We do know that individual sending institutions have responded to our feedback by addressing issues, but this knowledge has not been generalized to the community at large, so any new partner assessed is likely to have quality scores similar to the population average.

We identified several common conformance issues. The existence of these issues suggests that the onboarding process is not sufficient to ensure ongoing operational data quality, and that operational monitoring should be a critical part of the national interoperability infrastructure.

The issues most commonly identified include the following:

1. **Missing required element** (blank field), as required by the C32 specification. Common areas include laboratory result interpretations & reference ranges, encounter comments, allergy reactions and confirmation dates, and medication prescription numbers, dates, and sigs.

2. **Wrong terminology** (populated but with a code not from the specified terminology). Common errors concern units of measure and status values following record status (e.g., “complete”) rather than clinical status (e.g., “resolved”). Local codes and labels for encounters and procedures are often too vague or too organization-specific to be useful, though standard values are not required by the C32.

3. **Misplaced information**. Common errors include stipulation of dose magnitude in medication quantity (which should contain dispensed quantities) and inclusion of immunizations in the medication section. Less
clear are inclusion of radiology results in a results section, and laboratory qualifiers such as patient challenges or blood bank product identifiers.

These errors—clearly stipulated in the specification, yet commonly mis-implemented in incoming documents—confirm that the onboarding process is not sufficient to ensure operational conformance. There is overlap between this list and that provided by D’Amore et al. Our “misplaced data” is very similar to their “inappropriate organization,” and our “wrong terminology” is a large subset of their “terminology misuse.” We do not adopt the category “incorrect data,” as we feel the nature of the error can usually be specified. The “Heterogeneity” classifications from that paper are only represented here where an issue creates an error or a clarity issue—e.g., where, in their example, the sporadic reference to narrative text causes a required element to be missing.

Other errors common enough to require their own classification, include

4. **Terminology codes too general to be useful**; e.g., ICD-9 799.9 “Other unknown and unspecified cause of morbidity and mortality.” Usually, this indicates a free-text entry.

5. **Formatting errors**, including the concatenation of type codes (“home,” “work”) or other unidentified text with displayed text.

6. **Duplicated data** - elements that are either duplicated or are valid but so similar to other valid elements that they appear to be duplicates. This occurs more readily when elements are classified at a higher level in a terminology, where a more detailed classification might allow discrimination (see # 4).

We identified a smaller number of issues with the stylesheet used to present documents to clinicians. These findings are fundamentally internal to the VA; we share them in case others have similar issues.

1. Patient names may be repeated or concatenated with aliases.
2. Blood pressure readings are concatenated for a conventional “Systolic/Diastolic” display, but units are omitted.
3. The transform presents scientific notation in an unconventional and potentially ambiguous fashion. Most of the numbers in figure 4 are presented without symbols to orient the reader to the fact that this is scientific notation: “0.4 10” could easily be read as “0.410.” (This presentation issue is in addition to the sender error of not sending the UCUM unit—in the case of monocytes, “%”—making it difficult to compare, e.g., monocyte and granulocyte measurements, as well as the sender error of pairing “count” test names with percentage result values.)

| #BASO | -- | 0.4% | Normal | 0.3 - 1.0 |
| #LYMPH | -- | 1.10 | Low | 1.5 - 2.9 |
| #MONO | -- | 0.410 | Normal | 0.3 - 0.8 |
| #GRAN | -- | 4.010 | Normal | 2.8 - 4.8 |
| #EOSIN | -- | 0.110 | Normal | 0.00 - 0.20 |
| #BAFO | -- | 0.010 | Normal | 0.00 - 0.10 |
| Erythrocytes[/volume] in Blood by Automated count | -- | 4.56 x 10 | Normal | 4.00 - 6.00 |

**Figure 4**: Examples of ambiguous scientific notation in the lab section of a CDA document

The most important findings are those related to specification ambiguity. These issues cannot be solved by testing and feedback loops internal to any organization; they must be solved by consensus within the interoperability community. Given the timeframe necessary to modify a standard, interim measures could be taken within interoperability communities (e.g., Healtheway) by codifying rules that we expect or recommend be adopted by HL7 Structured Documents Workgroup (SD) or C-CDA. This approach has informed the adoption of the “Bridge c32”—a specification used by Healtheway to reduce optionality and ambiguity within its community of participants while
at the same time increasing the richness of the data—as well as the development of the C-CDA. It should be extended to leverage the collective experience of all participants.

These issues include questions about the validity of included data, how to format information consistently, and several questions regarding heterogeneity in the results section.

Questions about the validity of data:
1. What is the temporal scope of a summary document? How far into the past should it reach, and how does this rule vary per clinical topic (e.g., medications, presumably limited, vs. allergies, presumably without limit).
2. May a summary document include data from other sources? From the requestor?
3. If data has been corrected, how is this indicated?
4. For Health Information Exchanges, should the custodian be the HIE or the originating facility or the system that captured the information (e.g., ADT)? A clinician may be more interested in the facility. Documenting data provenance is an area where further specification is needed.
5. How specific should textual information be? Can providers be listed as "historical provider"; or can sourced institutions be listed as abbreviations?
6. Must medications include specific information (e.g., dose), or is the clinical drug alone sufficient?
7. How should fields that are not applicable be represented? For example, in a record of "no known allergies," how should the allergy reaction be consistently represented? As blanks? As "not applicable"?

Questions about the C32 results section:
1. How should result sections support different kinds of results—hematology and chemistry tests, microbiology tests, radiology reports, etc.?
2. How should ambiguous kinds of acts be listed? Is an x-ray a procedure or a result? What guidelines should be given to implementers regarding what goes into the Results section vs. the Procedures section?
3. How should sparsely populated result qualifiers be represented? E.g., patient challenges, blood bank products, method used, etc.
4. Should reports and clinical notes (e.g., radiology and surgery reports, progress notes, consult notes, discharge summaries, H&Ps) be included in a health summary?
5. If a result itself is an interpretation (e.g., “present”), should there be a redundant value in the interpretation field? An explicit null (e.g., “Not applicable”)? A blank?

Questions about the display text of data:
1. Should a coded element display the preferred term from the code system or a local or manually typed term? Should a code always be represented by the same term in a document? In a facility?
2. When a data element is described with more than one representation (a standard code, a translation code, and an original text), which text should be displayed to a clinician?
3. How should LOINC names be presented? The formal names are often too detailed for clinical use and the short names too abbreviated to be easily understood. Abbreviations affect all sections (e.g., ‘DMII WO CMP NT ST UNCNTR’ for a problem name, ‘O’ for an Encounter type).
4. Is there a problem with appending a code to a coded concept display term (e.g., appending the ICD9 code after the diagnosis name)? Should it be encouraged?
5. Should empty sections be allowed or prohibited? Does an empty section constitute useful information?
6. How can problem or lab lists be kept to a manageable length? In general, how can the relevant data be made more clear to the receiver?
7. Should the community enforce that the information in the narrative blocks and structured entries be equivalent? If a sender expects the narrative text, as the “attested” data, to be understood, whereas the receiver constructs the human display from the structured data, how can we know that they are semantically
equivalent? Should there be more prescriptive recommendation that the clinician display always be derived from the narrative blocks?

As we consider these questions, we observe that some of them can, once answered by the consensus of stakeholders, be implemented as verifiable rules in the evolving standard. Classification of laboratory information, for instance, should be done once, codified, and implemented. Other questions, however, will require a more nuanced response. Determination of the appropriate temporal or clinical scope of a summary, for instance, will likely be the result of an ongoing discussion among practicing clinicians. The solution to such a question is unlikely to be encoded in a conformance rule, at least in the near future. Instead, practicing clinicians will need a feedback channel to provide insight to document senders. Whether such a loop leverages the expertise of the clinician requesting the document in the first place or that of a dedicated review team remains to be seen, but either way, it should minimize the risk of imposing a burden on clinicians by supporting feedback as terse or detailed as the clinician feels is warranted.

Conclusion

In this paper, we share the VA experience with data quality surveillance of eHealth Exchange participants C32 documents. This surveillance is conducted with real patient data and produces several scores, including data richness, data ‘correctness’, and semantic level.

We offer five recommendations from this experience.

First, we have identified common issues that seem to result from ambiguity in the C32 specification. When the community has moved largely to the C-CDA format, we expect these issues will continue to cause problems, to a lesser degree, but we hope that our experience provides some of the data needed to ameliorate those issues.

Second, we show the importance of continuous surveillance of data quality in a production environment after the certification and onboarding phases are completed. It may be appropriate that certification be a necessary condition for participation in Healtheway, but it is not sufficient to guarantee quality. Each participating organization should consider a C-CDA quality monitor running within their firewall.

Third, we demonstrate that automation can enhance the understanding of data quality, and we support continued investment in tooling for this purpose. While we have seen enough documents to have confidence in our qualitative judgments, we look forward to having the ability to monitor these issues over larger samples, to do so in near-real time, and to use this data in operational feedback to improve the quality of care. These automated tools would come with a set shared public rules that each organization can implement in their environment to continuously assess and alert about the data quality index of the information exchanged. Interoperability is enhanced only when all exchange participants cooperate in monitoring and improvement. One organization cannot do it alone.

Fourth, we have identified several items that may not be amenable to automated testing. Processes for monitoring these issues will require human intervention. Initially, these processes can be conducted based on sampling, but methods should be assessed for capturing operational feedback.

Finally, we recommend that the efforts to identify ambiguities be centralized to support continuous improvement across the Healtheway eHealth Exchange community. As the questions we raise are answered and built into the specification, it will become incrementally more useful, but the list will continue to grow for some time to come, and a governance mechanism will be needed to manage their resolution in ways that continue to support interoperability. This mechanism should support reporting of issues, consensus on solutions, methods for sharing examples, and channels for advocating solutions to standards development groups.

References


Towards data integration automation for the French rare disease registry

Meriem Maaroufi, MSc1,2, Rémy Choquet, PhD1,2, Paul Landais, MD, PhD1,3, Marie-Christine Jaulent, PhD2

1 Banque Nationale de Données Maladies Rares, Hôpital Necker Enfants Malades, Assistance Publique des Hôpitaux de Paris, Paris, France; 2 INSERM, U1142, LIMICS, Paris, France; 3 Montpellier University, EA2415 & BESPIM, University Hospital, Nîmes, France

Abstract

Building a medical registry upon an existing infrastructure and rooted practices is not an easy task. It is the case for the BNDMR project, the French rare disease registry, that aims to collect administrative and medical data of rare disease patients seen in different hospitals. To avoid duplicating data entry for health professionals, the project plans to deploy connectors with the existing systems to automatically retrieve data. Given the data heterogeneity and the large number of source systems, the automation of connectors creation is required. In this context, we propose a methodology that optimizes the use of existing alignment approaches in the data integration processes. The generated mappings are formalized in exploitable mapping expressions. Following this methodology, a process has been experimented on specific data types of a source system: Boolean and predefined lists. As a result, effectiveness of the used alignment approach has been enhanced and more good mappings have been detected. Nonetheless, further improvements could be done to deal with the semantic issue and process other data types.

Introduction

Data analysis is important for health care systems optimization. Studying medical administrative data helps institutions to take right decisions toward health policy so as to lower costs and improve healthcare quality and access. In this context, the French health ministry is funding the BNDMR project (Banque Nationale de Données Maladies Rares) in order to build a medical data warehouse for rare diseases.

Currently, the French rare diseases (RD) care network is based on 131 rare disease centers of excellence. Each center is composed of one or more healthcare RD units, called sites, located in hospital structures that are known for their expertise and specific skills in managing and treating particular rare diseases.

In their daily activity, professionals from those sites use different systems to enter patients’ data such as administrative data, biological/imagery results and reports in the hospital information system; disease specific data in national or international registries; genetic or biological data in specific applications; customized data in local data bases, etc.

In order to avoid duplicating data entry for health professionals, the BNDMR will propose the deployment of connectors to retrieve data from the different source systems and allow interoperability at a national level.

The first step to ensure interoperability was to define a minimum data set built on common information to all rare diseases. This minimum data set was built and standardized following a methodology. The next step to collect data from all the different systems used by health professionals is to align the data schemas of source systems with the target data schema, the BNDMR minimum data set, in order to detect the mappings linking data elements. To manually align the data schemas is a tedious task. Even though we are familiar with the target data schema and its sixty data elements, there is a huge number of potential source systems, about 500 systems, with heterogeneous datasets that are more or less voluminous. The automation of this task becomes necessary.

The literature describes many automated alignment tools based on different approaches. Those tools have been used in different contexts to discover mappings between heterogeneous data sources e.g. BiobankConnect. Obtained results are promising and reliable but they are not sufficient to proceed directly to data exchange. An important consideration is missing: the values of data elements. If the value element level is not taken into account, first we can miss some mappings that exist at that level, and second, we cannot define a concrete mapping between data elements that can be directly implemented in the data integration process.

In this context, we previously proposed a mapping formalization in order to take into account the value element level that can be used to generate some parts of the program in charge of data import or export. In this paper we describe how this formalization contributes in a whole data integration process. Then, we present a concrete application by aligning a sub dataset of a data source schema, CEMARA, an academic rare disease data base, with the target data schema, the minimum data set of the BNDMR.
Background and proposal

Authors of different comparative studies on alignment or matching approaches agree to say that the efficiency of those approaches depends on data types and the application context. Furthermore, the aim of those tools is to align two sets of concepts (concepts of a thesaurus or an ontology). Using a certain measure, they detect pairs of concepts and assign them a similarity level. The obtained results are mappings in this format: \([C_1, C_2, s(C_1, C_2)]\), e.g. \([\text{Cystinosis}, \text{Cystin}, 0.8]\).

These results are not sufficient to define a complete data integration process. In fact, value elements are only considered when the alignment tools use instance-based approaches, but they are not integrated in the mappings definitions, they are only used to detect mappings between data elements. Data elements from the different schemas are often not coded the same way. Data transformation will be needed in order to complete the harmonization and proceed to data import. Data transformation includes: aligning lists of data values (thesaurus, specific lists...) and other operations like arithmetic operations to harmonize units or simple concatenation.

Rahm and Bernstein quickly mentioned some important issues, they pointed out the importance of “mapping expressions” to specify how the elements of two schemas are related. They also underlined the fact that, in practice, the criteria matching the elements are based on heuristics and are not mathematically precise to be used when implementing the match. On the basis of those considerations, it is necessary to: (i) optimize the use of the different alignment approaches in the data integration process while taking into account data types; (ii) generate well-defined mappings, integrating a value element level that can be directly exploitable.

(i) Integrate schema alignment approaches in a whole process

The alignment algorithms intervenes in the second step of a four phased data integration process:

- **First phase: data selection**
  
  As depicted in most studies, some alignment approaches can be considered as appropriate or inappropriate depending on data types. This phase will allow data selection according to the data types that will be addressed: numeric types, string and text types, Boolean type, lists and enumerations, etc. Data selection can be carried further by considering some constraints on data like strings length, bounds and units of numeric types, or source terminologies of some enumerations. It is also important to specify the context and domain of study in order to adapt the external resources that will be used: dictionaries and terminologies.

- **Second phase: mappings detection**
  
  In this phase, automatic alignment tools will be used to detect mappings in the way that have been designed for the selected data. This can be done by applying directly one alignment approach or by defining a complex path that will be taken by the data. External resources can be used to semantically enrich the data schemas and, for example, be able to map synonyms.

- **Third phase: mappings validation**

  Validation is generally a human validation involving persons familiar with the source and target data schemas.

- **Fourth phase: code generation**

  Once the mappings validated, they are integrated in the program in charge of data transformation and exchange. This implies that mapping expressions have to be readable by programming languages.

(ii) Mapping formalization

As proposed in our previous publication, a mapping from a source schema \(S\) to a target schema \(T\) can be characterized by the triplet \([E^S_i - E^T_j ; e^S_{ik} - e^T_{jl} ; r]\):

- A binary relation \(E^S_i - E^T_j\) between a source data element and a target data element. A data element is a unit of data identified by a name having a definition and that can be permissive to only some values.
- A binary relation \(e^S_{ik} - e^T_{jl}\) between a source value element of \(E^S_i\) and a target value element of \(E^T_j\). A value element is a value out of a set of permissible values pertaining to a data element.
- A rule \(r\) expressed in the “ if … then …” format defining the exact relation linking the different data and value elements.

This formalization is not a bijection; it links the source elements to the target elements and the reverse path is not always true.

Example: \(\{\text{“Coming CPC”} \rightarrow \text{“Patient addressed by”}; \text{“Y”} \rightarrow \text{“CPC”}; \text{if “Coming CPC” = “Y” than “Patient addressed by” = “CPC”}\}\)
Experimentation

We experimented our methodology by aligning a sub dataset of our first data source schema, CEMARA, with the target data schema, the minimum data set of the BNDMR.

- Data selection

The process we defined is able to detect simple mappings between source and target elements from Boolean and predefined lists data. Among the 72 data elements from CEMARA schema, 39 are Boolean data and 15 are predefined lists. As of the BNDMR schema, it contains 15 Boolean and 16 predefined lists among its 62 data elements.

Since the Boolean value elements (“true” and “false”) are not enough informative to infer mappings, we settled for considering only the data element level. For instance, we will try to map the data element’s label “the patient is deceased” and we will not do so for the value elements “true” or “false”. However the value elements will be integrated in the formalization of mappings into the rules during the final phase.

On the other side, considering the value element level becomes more important than considering the data element level when we are dealing with predefined lists. Furthermore, value elements from two lists can be mapped even though their respective data elements are not semantically similar.

E. g. the source data element “act type” and the target data element “participant profession” are two different concepts however their respective value elements, “nurse intervention” and “nurse” are closely linked.

The number of elements to map moves from 15 data elements to 86 value elements for the source CEMARA and from 16 data elements to 106 value elements for the target BNDMR.

- Mappings detection

A linguistic approach is used to operate a cross alignment between the different types of elements. This linguistic approach is based first on a semantic enhancement of both source and target data schemas and second on a lexical matching between the elements using the Levenshtein algorithm (also called the Levenshtein distance used to measure the differences between two sequences using the number of characters’ deletions, insertions and substitutions)(8). For the semantic enhancement, synonyms and English translations have been assigned to the elements.

OnAGUI⁸, a tool for aligning ontologies integrating different algorithms, has been used to apply the Levenshtein algorithm on the source and target schemas that have been translated in OWL format.

The same information can be structured differently from one schema to another. Cross alignment between different typed data is necessary. Four alignments have been done, operating either at the same level (data element or value element level) or at different levels (both data element and value element levels) depending on the type of data that we are aligning, as explained in the data selection phase (Figure 1).

![Figure 1. Alignment process: Cross alignment for Boolean data and predefined lists](image)

- Mappings validation

For our experimentation validation, a person familiar with both source and target databases has done a manual alignment between the different elements to establish a gold standard. Two experts of both databases validated this alignment. This liable set of mappings allowed the validation of mappings that have been detected by the defined process.
In routine use, a simple selection of the valid mappings would have been sufficient for this phase.

- Code generation

The table below (Table 1) shows that four simple types of rules will be generated by this process, linking Boolean value elements (true and false) and lists value elements.

**Table 1. Generated mappings expressions**

<table>
<thead>
<tr>
<th>Source elements</th>
<th>Target elements</th>
<th>Resulting mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data type</td>
<td>Element level</td>
<td>Data type</td>
</tr>
<tr>
<td>Matching 1</td>
<td>boolean DE*</td>
<td>boolean DE</td>
</tr>
<tr>
<td>Matching 2</td>
<td>list VE*</td>
<td>boolean DE</td>
</tr>
<tr>
<td>Matching 3</td>
<td>boolean DE</td>
<td>list VE</td>
</tr>
<tr>
<td>Matching 4</td>
<td>list VE</td>
<td>list VE</td>
</tr>
</tbody>
</table>

* DE: data element, VE: value element

Below some examples of the generated mapping expressions:

- If PropLink=propositus [source] then Propositus=true [target]
- If ConfCyto=true [source] then ConfirmationMode=cytogenetic [target]

**Results**

The similarity threshold for the Levenshtein distance has been set to 0.9 to contain the number of false positive mappings. Counting is relative to the target elements: How many target data elements and value elements have been mapped? How many have been correctly mapped or not mapped?

In Matching 1 and Matching 2, the algorithm is used to search mappings with source elements for the 16 Boolean target data elements. In Matching 3 and Matching 4, the algorithm is used to search mappings with source elements for the 106 target value elements building the predefined lists of some data elements (Table 2).

**Table 2. Alignment results**

<table>
<thead>
<tr>
<th></th>
<th>Good mappings</th>
<th>Wrong mappings</th>
<th>Missed mappings</th>
<th>Good non mappings</th>
<th>Total</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP*</td>
<td>FP*</td>
<td>FN*</td>
<td>TN*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matching 1</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>0,43</td>
<td>1,00</td>
</tr>
<tr>
<td>Matching 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>16</td>
<td>1,00</td>
<td>1,00</td>
</tr>
<tr>
<td>Matching 3</td>
<td>22</td>
<td>4</td>
<td>4</td>
<td>76</td>
<td>106</td>
<td>0,85</td>
<td>0,85</td>
</tr>
<tr>
<td>Matching 4</td>
<td>35</td>
<td>8</td>
<td>25</td>
<td>38</td>
<td>106</td>
<td>0,58</td>
<td>0,81</td>
</tr>
</tbody>
</table>

* TP: true positives, FP: false positives, FN: false negatives, TN: true negatives

For a comparative evaluation, we ran the Levenshtein algorithm on the complete source and target data schemas as a reference test. We kept the same threshold of 0.9 of similarity. Fifteen mappings have been detected linking 14 CEMARA data elements to 14 BNDMR data elements. Among the 15 mappings, 5 were false mappings, e.g. Diagnostic status – Patient status. Operating on a data element level can be misleading especially when dealing with predefined lists. The 10 good mappings are distributed as shown below (Table 3).
Table 3. Comparative evaluation: number of good mappings for the reference test and the experimentation

<table>
<thead>
<tr>
<th></th>
<th>Reference Test</th>
<th>Experimentation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching 1</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Matching 2</td>
<td>0</td>
<td>1</td>
<td>Not the same total number of elements. In the reference test, only data elements are considered</td>
</tr>
<tr>
<td>Matching 3</td>
<td>1 (2 missed)</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Matching 4</td>
<td>6 (2 missed)</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

Besides the missed mappings, mappings that have been detected by the reference test are linking data elements and not value elements. Thus, seven of the ten “good” mappings are not precise enough to be considered as good mappings.

Discussion

To summarize, the proposed method is based on a new way to formalize mappings with rules specifying the exact relationship between source and target data elements and their values. The alignment process is designed to suit the data types chosen as inputs in order to generate well-defined type of rules as output. In our experimentation we studied two data types: Boolean and predefined lists, we operate a cross-alignment at both data element and value element levels and detected four types of rules defining the mappings.

The comparative evaluation shows that using the same data schemas (CEMARA and BNDMR) and the same alignment tool (OnAGUI) as the reference test, more mappings could have been detected when using the proposed methodology. Further, value elements form integral part of the mapping expressions. Moreover, results show that globally we obtained good recall and precision. Wrong mappings (false positives) that move away the precision from 1 can be explained by the generic nature of some value elements, e.g. “other”. Their occurrence in different list data elements explains their wrong mapping to similar elements several times. Such cases can be processed differently in the alignment process by linking them to their data element or siblings value elements. However, precision remains acceptable, greater than 0.8. The recall is more related to semantic issues. Some mappings have not been detected (false negatives) this is mainly due to the chosen alignment approach that is based on a simple string similarity metric. Moreover, and despite the efforts made to build and enrich the schemas, the algorithm does not use at the moment an external semantic resource to search more synonyms or vocabulary variants, like the difference between “medically assisted procreation” and “medical assistance to procreation” and thus improve on the recall.

If the data integration process were totally automated with no human intervention to validate or reject the proposed mappings, the registry data quality will be negatively affected. A bad recall would mean a non-efficient alignment with missing mappings and therefore missing data. A bad precision would mean wrong mappings and therefore importing false information in the target data base. The priority in our work is to maintain the precision as high as possible while trying to improve the recall making the task easier for the user which validation remains essential to not affect the data base quality and maintain the reliability of the ensuing studies. Thus, among the next evolutions of our work, an automatic semantic enrichment will be integrated to the process. The first step would be to request the MeSH terminology to retrieve data elements and value elements synonyms to catch some undetected mappings. It would also be interesting to consider other data types and define the suitable alignment process to detect their mappings. E.g. in order to integrate some measures like the patient weight and height, numeric data alignment may be addressed with taking into account the unit conversion. And finally, a global validation will be conducted using other source systems.

Conclusion

The proposed methodology aims to optimize the use of existing alignment approaches and to limit human intervention in the data integration process. Our work takes part of a global effort for the healthcare systems interoperability. Data schemas alignment can be an alternative to address the lack of standardization in the existing systems or as a complement if different standards are implemented.

Acknowledgments

We acknowledge the BNDMR project team for their comments. This work was funded by the French Ministry of Health.
References

Diagnostic Characteristics of Patient Self-Assessment of Preoperative Cardiac Risk for Non-Cardiac Surgery - Foundations for Patient Driven Decision Support

Sharad Manaktala MD, PhD 1; Todd Rockwood PhD 4; Terrence J. Adam RPh MD, PhD1,2,3

1 Institute for Health Informatics; 2 Minneapolis VA Healthcare System; 3 College of Pharmacy; 4 Health Services Research and Policy, School of Public Health, University of Minnesota

Introduction
Surgery is a frequent health care intervention with an estimated 6 million non-cardiac surgical procedures performed every year in the United States, with progressive growth in the number of procedures noted each subsequent year(1). Preoperative medical evaluations have been a cornerstone of pre-surgical care planning and patient management. The clinical providers who carry out these pre-operative evaluations are guided by their medical knowledge, institutional policies and clinical practice guidelines. To address the challenge of standardizing the evaluation and management of perioperative cardiovascular complications of non-cardiac surgery, the American College of Cardiology/American Heart Association (ACC/AHA) issued perioperative guidelines for non-cardiac surgery in 2007 with a focused update on preoperative beta blockade in 2009(2), followed by the latest update in 2014 (3). These guidelines modified recommendations on the Revised Cardiac Risk Index (RCRI) (4,5) to highlight a set of conditions associated with higher post-operative morbidity and mortality for patients undergoing non-cardiac surgery.

As surgical interventions continue to evolve with increasingly complex and costly procedures, there is a critical need to improve the preoperative assessment to effectively identify clinical risk factors and manage existing co-morbidities. In the typical perioperative window, from one month before to one month after the surgical intervention, targeted risk mitigation can be implemented to reduce surgical complication risk. In the context of growing surgical work volumes, there is a paucity of well-trained primary providers and preoperative assessment clinics to address patient needs with ongoing growth in the number of preoperative assessments (6-8).

An important partner for the surgical team is the patient, who ultimately has the most at stake from the surgical benefits and potential complications (9). Unfortunately, few tools are available for patients to self-identify surgical risk and empower them to work in tandem with multi-disciplinary surgical teams. Such tools could help patients become better informed of their surgical risk and address an important knowledge gap since most patients have limited recall of the risks and benefits of surgical interventions after completing the pre-operative clinical workup (10,11). In addition to the potential educational and clinical benefits to patients, patient driven decision support tools could also be a cost-effective adjunct tool for surgical quality efforts including Accountable Care Organizations (ACO), medical homes, and other efforts to enhance clinical quality. Having effective patient driven clinical assessments can provide surgical providers a greater appreciation of risks prior to their planned procedures and facilitate optimized multi-specialty care delivery.

Decision support tools that offer patient self-assessment can potentially help mitigate time pressures in the typical 30-day preoperative window by adding important triage data to better identify the at-risk populations. To use patient-driven clinical information to its optimal capacity, it is important to establish its validity and reliability. The ability of patients to self-identify clinical risk factors has been previously established in many studies of chronic disease (12-17). There are several strategic advantages to having patients identify their own clinical risk factors, both to have patients better understand their own clinical conditions and risk of surgical complications, as well as potential cost savings with obtaining valid clinical information without the cost of provider mediation (11,18,19). However, tools for patient self-identification of clinical risk factors have not always crossed directly from large research studies into use in clinical care (9,17). Technologies like clinical decision support systems (CDSS) hold immense potential to enable patients to access higher tiers of the knowledge pyramid, thereby gaining from the collective wisdom of scientific literature.

This study aims to create and validate prediction rules to form the basis for a future patient generated risk prediction tools. These validated tools integrated into electronic and patient interfacing formats like personal health records, mobile and internet-based applications can be instrumental in promoting patient-centric health care and timely risk stratification of preoperative risk.
Methods

Experimental Design - A prospective, single center, hospital based observational study was performed to evaluate the diagnostic characteristics of patient self-reported health information, after approvals were received from the Institutional Review Board (IRB).

Site Description - The Minneapolis Veterans Administration Hospital is a major referral site within the VA system with care provision by multiple surgical specialties in the VISN 23 region. The VISN 23 clinical health care network serves more than 400,000 enrolled Veterans residing in the states of Iowa, Minnesota, Nebraska, North Dakota, South Dakota and portions of Illinois, Kansas, Missouri, Wisconsin and Wyoming. The preoperative medical clinic operated from the Minneapolis clinical site is a large preoperative medical site with 10 clinicians providing preoperative medical assessments.

Survey Design - To accurately assess and capture patient medical and surgical history, exercise tolerance, and cardiovascular risk perceptions; a survey instrument was iteratively developed in the form of a 25-point assessment tool. This instrument was developed by mapping the recommendations of American Heart Association (AHA) Guidelines for Pre-operative risk assessment of Cardiac complications of Non-Cardiac Surgery (2) The AHA recommends using the Revised Cardiac Risk Index (RCRI) (5) which stratifies cardiovascular complication risk based upon 6 key risk factors: history of coronary artery disease (CAD) or unstable angina, history of cerebrovascular accident, history of Congestive Heart Failure (CHF), diabetes mellitus requiring insulin for control, serum creatinine greater than 2mg/dL, and high risk surgical procedure. A questionnaire survey instrument was developed by modifying existing patient self-report instruments including: Modified Rose Questionnaire for IHD (20,21); Questionnaire to Verify Stroke Free Status (QVSFS) (22) for Stroke and Cerebrovascular accident (CVA); and Compendium of physical activity (23) to assess patient exercise capacity. Symptomatic questions were developed based on expert consensus when a suitable validated instrument was not available for risk factor assessment. After identifying each of the six revised cardiac risk index factors, a patient reported RCRI score was generated along with other patient reported health information such as cardiovascular and procedure history. In addition to objective cardiac risk factor assessment, questions were developed to identify patient cardiac risk perceptions on a graded scale. These results were scored with a mapping algorithm to identify the relative levels of risk perceived by patients while reflecting current clinical guidelines. The survey content was tailored for patient use with an estimated reading level of grade 7 by Flesch-Kincaid readability analysis to facilitate patient use. Questions on important surgical contraindications including certain high risk surgical preclusion criteria, pre-existing conditions, cardiovascular risk perception, exercise tolerance, and the elements of the 6 revised cardiac risk index risk factors were adapted for patient use and incorporated into the instrument.
**Survey Validity** - An intended referential was established along with the connotative meaning of each question to ensure that the respondents interpreted the questions as intended. The survey was developed while making broad consideration for the educational level of a wide range of participants but focused on those in the VA Medical Center, Minneapolis. The survey content was tailored for patient use with an estimated readability level < grade 8 on the Flesch-Kincaid Readability Scale (24,25), consistent with readability level of a high school graduate. The questionnaire was designed and assessed to avoid biased wording, double negatives and leading questions to check for response bias. Closed ended questions were framed around almost all questions with ordered and partially ordered responses for most questions, to allow a continuum of responses and to facilitate ease and accuracy of responses (26). Survey length was kept at less than 25 questions that were spread over 3 letter-sized pages. Survey completion times were recorded during pre-testing in a sub-set of participants to ensure that the respondents are able to complete the questionnaire within a reasonable time frame. Pre-testing on a small sub-set of participants revealed that all participants were able to complete the survey within 3 to 4 minutes. Concordance between patient and provider reported RCRI scoring was evaluated in a separate study (27).

**Data Collection** – Patients checking in at the preoperative medicine clinic were asked to complete the survey questionnaires while they waited to be seen by their clinicians. Patients at the preoperative medicine clinic completed a total of 500 surveys during the study period. Since the survey development process was iterative, only 401 patients, who completed the finalized survey version, were included in the study. Patients visiting the preoperative medicine clinic in the Department of General Medicine VAMC Minneapolis during the study period were included in the study. Provider data was retrospectively collected from chart review of provider visit notes, problem list, drug list and surgical records of corresponding patients from the VistA electronic medical record system. We excluded eighty-seven more patients that had missing or incomplete survey responses and for some cases incomplete or cancelled preoperative clinician evaluations. Thus, only the remaining 314 surveys and their corresponding clinician evaluations were used for comparison of RCRI scores, risk perception and validity analyses.

**Data Analysis** - Descriptive statistics were used to tabulate patient demographics, cardiac risk self-perceptions, and cardiovascular risk profiles using SAS version 9.3 (SAS Institute Cary, NC). We evaluated the association between provider perceptions of patient risk, provider generated RCRI score with patient subjective perception of risk and RCRI scores generated by patient provided survey information. We calculated Cronbach’s alpha to estimate the internal consistency of this survey.

**Results**

At the end of the study period we included study participants with ages ranging from 25 to 91, with an average age of 66 (+12.4) years for the study population. Typical for veteran populations (28) who constituted the majority of study participants, 9 out of 10 participants were males. Other details for patient demographic data is presented in table 2. The calculated standardized Cronbach’s alpha (0.93) indicated that the survey had excellent internal consistency.

<table>
<thead>
<tr>
<th>Males n (%)</th>
<th>Mean (+SD){Range}</th>
<th>25 - 44 years n (%)</th>
<th>45-64 years n (%)</th>
<th>65-80 years n (%)</th>
<th>&gt; 80 years n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>381 (95%)</td>
<td>66 years (+12.4) {25-91}</td>
<td>35 (9%)</td>
<td>195 (49%)</td>
<td>128 (32%)</td>
<td>43 (11%)</td>
</tr>
</tbody>
</table>

Table 1 - Gender and age distribution for study population

<table>
<thead>
<tr>
<th></th>
<th>Coronary Heart Disease</th>
<th>Congestive Heart Failure</th>
<th>Cerebrovascular Accident/Stroke</th>
<th>Diabetes requiring Insulin</th>
<th>Renal Insufficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (%)</td>
<td>85 (21%)</td>
<td>28 (7%)</td>
<td>24 (6%)</td>
<td>38 (9%)</td>
<td>17 (4%)</td>
</tr>
</tbody>
</table>

Table 2 – Prevalence of provider diagnosed risk factors in study population

We also calculated frequencies and prevalence of cardiovascular risk factors outlines in the RCRI using descriptive statistics. We noted that coronary heart disease was the most common risk factor with a prevalence of 21% (n = 85),
followed by Diabetes requiring insulin 9% (n = 38), with renal insufficiency observed as least prevalent at 4% (n = 17) (Table 2).

To evaluate concordance between RCRI scores generated by patient provided information and comprehensive chart review and provider assessment, we created a comparison matrix (Table 3).

<table>
<thead>
<tr>
<th>Patient RCRI</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>199</td>
<td>15</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>44</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>8</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SUM</td>
<td>213</td>
<td>68</td>
<td>20</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3 - Comparison Matrix of Patient Generated RCRI scores with Expert Provider Assessment {Absolute difference between provider generated and patient RCRI score: Green - Zero, Yellow - One, Red - More than one}

This RCRI comparison demonstrated complete agreement between the patient and provider scores in 261 out the total 314 surveyed patients, with some level of disagreement in the remainder. This observation illustrates the fact that patient identified risk factors and the generated RCRI scores were in 83% agreement. There were only a limited number of subjects (n = 8) for whom the absolute difference between patient driven and provider RCRI scores were greater than 1.

**Patient Self-Perceptions of Cardiovascular Risk**

To explore potential drivers of patient self-perception of cardiovascular risk, we evaluated the strength of association between patient self-perceived risk and gold standard physician diagnosed risk factors (Table 4). We noted that patients who had physician diagnosed CAD were 5 times more likely to perceive themselves as high risk for cardiovascular complications peri-operatively compared to patients who had a negative history for physician diagnosed CAD, despite the nearly same risk of other factors such as renal insufficiency. Similarly, patients with physician reported history of CABG were 8 times more likely to consider themselves as high risk for cardiovascular complications peri-operatively compared to those who had not had a CABG. The perceived risk association for stroke, heart failure, diabetes requiring insulin and renal insufficiency, were however, statistically insignificant.

**Validity of Patient Self-reported Health Information**

We also performed an analysis of the operating characteristics for validity of patient self-diagnosis versus gold standard clinician diagnosis for peri-operative risk factors. As seen in Table 5, all patient self-reported risk factors had a statistically significant overall accuracy of at least 0.93 compared to gold standard clinician evaluation. We also noted that all patient reported risk factors including coronary heart disease (CAD), cerebrovascular accident (CVA), diabetes, and exercise tolerance had acceptable sensitivity values of 0.77, 0.83, 0.87, and 0.86 respectively. However, all five patient reported risk factors had high specificities, Negative Predictive Values (NPV) and positive Likelihood Ratios (LR +).

**Clinician perception of patient risk for complications**

The validation study by Goldman et al concluded that all risk factors (CAD, CVA, diabetes requiring insulin, CHF, renal insufficiency and high risk type surgery) had an equivalent contribution to the risk of post-operative cardiovascular complications after non-cardiac surgery (4). However, we observed that certain patient risk factors influenced clinician perceived risk more than the others. For example, we used logistic regression to calculate odds ratios for evaluating the association between patient risk factors and provider perceptions of post-operative adverse
cardiac events. We observed that after adjusting for age, providers perceived patients who self-reported a history of CAD to be 9 times (95% C.I. 4.98-16.45) higher risk of having an adverse cardiac event after surgery as compared to patients who had no history of heart disease. History of chest pain, history of CHF, presence of cardiac pacemaker or defibrillator, high-risk type planned procedure were among the other significant driver of clinician risk perception for their patients. (Table - 6)

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Odds Ratio (95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician Diagnosed CAD</td>
<td>5.48* (2.51-11.95)</td>
</tr>
<tr>
<td>Physician Diagnosed CVA</td>
<td>2.6 (0.97-6.96)</td>
</tr>
<tr>
<td>Physician Diagnosed CHF</td>
<td>0.93 (0.2-4.2)</td>
</tr>
<tr>
<td>Physician Diagnosed Renal Insufficiency</td>
<td>Not significant</td>
</tr>
<tr>
<td>Diabetes requiring insulin (Self reported)</td>
<td>0.89 (0.25-3.16)</td>
</tr>
<tr>
<td>H/o of intra-cardiac device</td>
<td>1.84 (0.15-22.47)</td>
</tr>
<tr>
<td>H/o of CABG</td>
<td>8.53* (3.43-21.17)</td>
</tr>
<tr>
<td>H/o of Stress Test</td>
<td>0.89 (0.25-3.16)</td>
</tr>
</tbody>
</table>

Table 4 - What drives patient risk perception? Table shows association (odds ratio) between patients’ self-perception of risk and physician diagnosed risk factors adjusted for age (* Indicates p-value < 0.05)

<table>
<thead>
<tr>
<th>Patient Identified Risk Factor</th>
<th>Sensitivity (95% C.I.)</th>
<th>Specificity (95% C.I.)</th>
<th>PPV (95% C.I.)</th>
<th>NPV (95% C.I.)</th>
<th>LR+ (95% C.I.)</th>
<th>LR– (95% C.I.)</th>
<th>Overall Accuracy (95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ischemic Heart Disease</td>
<td>0.771 (0.66-0.85)</td>
<td>0.975 (0.94-0.98)</td>
<td>0.9 (0.79-0.95)</td>
<td>0.937 (0.9-0.96)</td>
<td>31.3 (14.08-69.85)</td>
<td>0.23 (0.15-0.36)</td>
<td>0.93 (0.9-0.95)</td>
</tr>
<tr>
<td>Congestive Heart Failure</td>
<td>0.6 (0.38-0.81)</td>
<td>0.969 (0.94-0.98)</td>
<td>0.579 (0.36-0.76)</td>
<td>0.972 (0.95-0.99)</td>
<td>19.6 (9.38-40.92)</td>
<td>0.412 (0.24-0.7)</td>
<td>0.94 (0.92-0.97)</td>
</tr>
<tr>
<td>Cerebrovascular Accident/Stroke</td>
<td>0.833 (0.6-0.94)</td>
<td>0.958 (0.92-0.97)</td>
<td>0.556 (0.36-0.78)</td>
<td>0.989 (0.96-0.99)</td>
<td>20.48 (11.33-37)</td>
<td>0.174 (0.06-0.48)</td>
<td>0.95 (0.92-0.97)</td>
</tr>
<tr>
<td>Diabetes Requiring Insulin</td>
<td>0.875 (0.69-0.95)</td>
<td>0.986 (0.96-0.99)</td>
<td>0.84 (0.65-0.93)</td>
<td>0.989 (0.96-0.99)</td>
<td>60.37 (22.55-161.59)</td>
<td>0.127 (0.04-0.36)</td>
<td>0.977 (0.96-0.99)</td>
</tr>
<tr>
<td>Serum Creatinine &gt;2mg/dL</td>
<td>0.357 (0.16-0.61)</td>
<td>0.972 (0.94-0.98)</td>
<td>0.385 (0.17-0.64)</td>
<td>0.969 (0.94-0.98)</td>
<td>12.81 (4.8-34.14)</td>
<td>0.661 (0.44-0.97)</td>
<td>0.945 (0.92-0.97)</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>0.867 (0.82-0.9)</td>
<td>0.64 (0.41-0.87)</td>
<td>0.976 (0.94-0.98)</td>
<td>0.123 (0.12-0.34)</td>
<td>2.31 (1.22-4.36)</td>
<td>0.212 (0.13-0.34)</td>
<td>0.855 (0.81-0.89)</td>
</tr>
</tbody>
</table>

Table 5 Diagnostic characteristics of Patient identified risk factors against gold standard expert clinician diagnosis (all p-values < 0.05)
Table 6 - What drives clinician risk perception? Association between provider perception for risk of cardiovascular complications and specific risk factors adjusted for age) {All number are odds ratios} [Questions for ψ – CAD, φ – Stroke, * Indicates p-value < 0.05]

Discussion
The results of this study provide key insights on the association between patient self-reported health information and expert clinician diagnosis of preoperative cardiovascular risk for patients undergoing non-cardiac surgery.

We investigated the concordance between RCR scores generated by gold standard provider assessment and patient self-identified risk factors, and generated RCRI scores were in near perfect (83%) agreement. We also concluded that patient reported coronary heart disease, congestive heart failure, presence of pacemaker or defibrillator, were perceived by provider to have higher risk for cardiovascular complications.

Upon examining the validity of patient reported health information, we concluded that patient self-reported RCRI risk factors (Ischemic heart disease, congestive heart failure, cerebrovascular accident, diabetes requiring insulin, and renal insufficiency) had a good and statistically significant overall accuracy, high specificity and negative predictive values compared to gold standard clinician evaluation. This underscores the fact that detailed pre-operative testing, evaluation and resources can be reserved for patients who report to have these risk factors, potentially saving on hospital costs and clinician time and avoiding risk of unnecessary testing and assessment which may create iatrogenic risk of otherwise low risk patients.

The results from our study agree with a large body of scientific literature suggests that patient self-reported health information is a valid resource of information and holds immense potential for improving the provision of healthcare (12,13,17,18,20,29,30). We acknowledge that there are number of limitations of this study. The study sample size is

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>OR (95 % C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.07 (1.04-1.1)*</td>
</tr>
<tr>
<td>Patient reported history of CAD ψ</td>
<td>9.05 (4.98-16.45)*</td>
</tr>
<tr>
<td>Does walking at a normal pace give you chest pain?ψ</td>
<td>24.28 (5.63-104.57)*</td>
</tr>
<tr>
<td>Does walking fast uphill give you chest pain? ψ</td>
<td>12.73 (4.45-36.34)*</td>
</tr>
<tr>
<td>Do you ever have any pain or discomfort in your chest? ψ</td>
<td>4.18 (2.18-8.03)*</td>
</tr>
<tr>
<td>Have you ever been told by a physician that you have had a stroke, mini-stroke, or transient ischemic attack (TIA)?ψ</td>
<td>3.51 (1.54-8.01)*</td>
</tr>
<tr>
<td>Have you ever had sudden painless weakness on one side of your body? φ</td>
<td>2.2 (0.66-7.37)</td>
</tr>
<tr>
<td>Have you ever had sudden numbness or a dead feeling on one side of your body? φ</td>
<td>2.33 (1.01-5.36)*</td>
</tr>
<tr>
<td>Have you ever had sudden painless loss of vision in one or both eyes? φ</td>
<td>2.13 (0.6-6.72)</td>
</tr>
<tr>
<td>Have you ever suddenly lost the ability to understand what other people are saying? φ</td>
<td>1.87 (0.47-7.35)</td>
</tr>
<tr>
<td>Have you ever suddenly lost the ability to speak or write? φ</td>
<td>0.39 (0.04-3.38)</td>
</tr>
<tr>
<td>Have you ever been told by a doctor that you have congestive heart failure (CHF)?</td>
<td>10.74 (4.03-28.62)*</td>
</tr>
<tr>
<td>Do you have leg swelling due to fluid retention?</td>
<td>2.95 (1.44-5.98)*</td>
</tr>
<tr>
<td>Have you ever been told by a physician about or seen your lab results showing blood creatinine level greater than 2 mg/dL?</td>
<td>4.93 (1.55-15.65)*</td>
</tr>
<tr>
<td>Have you ever been on dialysis?</td>
<td>2.4 (0.53-11.18)</td>
</tr>
<tr>
<td>Do you use insulin for treatment of diabetes?</td>
<td>6.15 (2.64-14.31)*</td>
</tr>
<tr>
<td>Pacemaker or ICD (Defibrillator) placement</td>
<td>14.7 (2.24-97.5)*</td>
</tr>
<tr>
<td>Heart bypass Surgery (CABG)</td>
<td>3.84 (1.73-8.52)*</td>
</tr>
<tr>
<td>Cardiac Stress test (treadmill test)?</td>
<td>1.79 (1.01-3.18)*</td>
</tr>
<tr>
<td>Insufficient Exercise Capacity</td>
<td>2.33 (1.1-4.5)*</td>
</tr>
<tr>
<td>Patient Self Perception of Risk</td>
<td>3.6 (1.89-6.87)*</td>
</tr>
<tr>
<td>High Risk Procedures</td>
<td>10.55 (3.03-36.68)*</td>
</tr>
<tr>
<td>Intermediate Risk Procedures</td>
<td>1.69 (0.84-3.3)</td>
</tr>
</tbody>
</table>
relatively modest, which makes it difficult to ensure that the patient self-report of cardiac risk factors and self-perception data is a true representation of Veterans Administration preoperative patients. The data collected in the study focused on a regional Veterans Administration pre-operative patient population and the data was generated from a single clinical practice site. It is unclear if these results could be extended to non-VA clinical sites given the predominant male distribution of the sample, which mandates further study. In addition, since this was a self-administered survey, the authors could not assess behavioral factors like patient denial or over-report for subjective patient health perceptions, but the availability of provider assessment data on the patient risk factors allowed for within patient comparisons. The authors also limited their scope to the assessment for perioperative cardiovascular complications to minimize bias. Thus, patient and provider perceptions of post-operative risk that may also be driven by other risk factors and pathologies like gender, age, race, and other co-morbidities.

In conclusion, we developed a patient self-report survey instrument that had excellent internal consistency. We also concluded that patient self-report is a valid resource for obtaining health information for preoperative cardiovascular evaluation. Our future goals include validating our findings through a study extending to a larger, non-veteran population; develop and evaluate other risk factors (like respiratory, neurological and medication management components) from patient driven data acquisition and management of perioperative risk, empowering patients in playing the pivotal role in their health care decisions.
References


Expanding a First-Order Logic Mitigation Framework to Handle Multimorbid Patient Preferences

Martin Michalowski, Ph.D\(^1\), Szymon Wilk, Ph.D\(^2\), Daniela Rosu, Ph.D\(^3\), Mounira Kezadri, Ph.D\(^3\), Wojtek Michalowski, Ph.D\(^3\), Marc Carrier, MD\(^4\)

\(^1\)Adventium Labs, Minneapolis, MN, USA; \(^2\)Poznan University of Technology, Poland; \(^3\)University of Ottawa, Canada; \(^4\)Ottawa Hospital Research Institute, Canada

Abstract

The increasing prevalence of multimorbidity is a challenge for physicians who have to manage a constantly growing number of patients with simultaneous diseases. Adding to this challenge is the need to incorporate patient preferences as key components of the care process, thanks in part to the emergence of personalized and participatory medicine. In our previous work we proposed a framework employing first order logic to represent clinical practice guidelines (CPGs) and to mitigate possible adverse interactions when concurrently applying multiple CPGs to a multimorbid patient. In this paper, we describe extensions to our methodological framework that (1) broaden our definition of revision operators to support required and desired types of revisions defined in secondary knowledge sources, and (2) expand the mitigation algorithm to apply revisions based on their type. We illustrate the capabilities of the expanded framework using a clinical case study of a multimorbid patient with stable cardiac artery disease who suffers a sudden onset of deep vein thrombosis.

Introduction

Mitigating clinical practice guidelines (CPGs) for a multimorbid patient is identified as a crucial step in the adoption of CPGs at the point of care\(^3\). The key problem lies in identifying and mitigating adverse interactions between guidelines given a specific multimorbid patient encounter. In addition to the CPGs, secondary knowledge applicable to both the multimorbidities and the patient state, and patient preferences must be used to propose a consistent therapy for the particular patient encounter. In previous work we proposed a logic-based approach to the mitigation problem that uses first-order logic (FOL) as the formalism for representing CPGs, patient information, and secondary medical knowledge in the form of revision operators\(^3,4\). With the emergence of personalized and participatory medicine\(^5\), patient preferences have become important components of the care process\(^6,7\).

Representations for CPGs, secondary medical knowledge, and patient information have been studied and formalized\(^8,9\), including in our own previous research\(^3,4,10,11\). Yet up to this point preferences were applied to decisions made together by the physician and patient (shared decision making) or by the patient (informed decision making). Furthermore these preference were represented at a very high level and not considered alongside revisions made to CPGs, making it difficult to identify and mitigate adverse interactions introduced by these preferences. Therefore, a multimorbid patient’s preferences must be represented formally so they can be part of the reasoning needed to mitigate concurrently applied CPGs.

In this paper we show how our FOL-based mitigation framework is flexible enough to support patient preferences and we describe how these preferences are represented in a format that is compatible with our FOL-based CPG formalism. Specifically, we broaden our previous definition of an operator and define revisions described in secondary medical knowledge and in preferences as two classes of operators. We also introduce a slight modification to our mitigation algorithm to apply as many preferences as possible while maintaining the consistency of the proposed therapy. Finally, we use a clinical scenario with a multimorbid patient who suffers from deep vein thrombosis (DVT) and a pulmonary embolism (PE) while being treated for stable cardiac artery disease (SCAD) and who has stated preferences for novel oral anticoagulants versus low molecular weight heparin and warfarin.

Supporting Preferences in a First-Order Logic Mitigation Framework

In this section, we describe the extensions to our first-order logic-based mitigation framework to support patient preferences. Our most recent published work covered our mitigation framework’s transition to FOL\(^1\), including a case study of a patient managed for type 2 diabetes and an onset of severe rheumatoid arthritis\(^4\), and our extension of the framework to support a broader set of CPGs (such as those that include parallel paths) and more complex secondary medical knowledge\(^12\). The extensions presented in this work serve two purposes: (1) to represent patient preferences and (2) to apply preferences during the mitigation process rather than as a post-processing step. These
extensions demonstrate the flexibility of our mitigation framework in supporting an additional form of secondary knowledge (patient preferences). To ground the extensions described below, we briefly reintroduce those framework components that are affected by or extended to support patient preferences. These include the combined mitigation theory $\mathcal{D}_{\text{comb}}$, revision operators, and the mitigation algorithm.

FOL is a widely used formal system for representing and reasoning about knowledge\textsuperscript{13}. This knowledge is represented as a logical theory $\mathcal{D}$, which is a collection of sentences in a first-order language $L$ defined over vocabulary $V$. $L$ consists of logical symbols (quantifiers, connectives, variables, and logical constants) and a finite number of non-logical symbols (predicates and functions) from $V$. The former have a fixed meaning in any FOL language, the latter are domain-dependent. The meaning of the non-logical symbols in $L$ is given by an interpretation $I$. If $I$ satisfies all sentences in $\mathcal{D}$, then it is called a model for $\mathcal{D}$. Theory $\mathcal{D}$ is consistent, iff there exists at least one model of this theory, and consistency is checked using theorem proving techniques. If $\mathcal{D}$ is consistent, it is possible to find models for this theory using model finding techniques, and to check for implications (logical consequences) of this theory through entailment\textsuperscript{3}.

In our work we assume a specific patient encounter is represented as the combined theory $\mathcal{D}_{\text{comb}}$. $\mathcal{D}_{\text{comb}}$ uses our defined vocabulary (see\textsuperscript{3} for its description) and is defined formally as the triple:

$$\mathcal{D}_{\text{comb}} = (\mathcal{D}_{\text{common}}, \mathcal{D}_{\text{cpg}}, \mathcal{D}_{\text{pi}}),$$

where $\mathcal{D}_{\text{common}}$ is a theory that axiomatizes the universal characteristics of CPGs as part of a FOL representation. It is the common (shared and reusable) component of all mitigation theories. $\mathcal{D}_{\text{cpg}}$ is a union of theories, each theory representing a single CPG that is being applied to a multimorbid patient. And $\mathcal{D}_{\text{pi}}$ is the theory that describes available patient information. It contains FOL sentences representing patient data, including results of tests and examinations, or indicates already prescribed therapies and procedures.

In our mitigation framework, all secondary knowledge is captured in revision operators. In our recent work\textsuperscript{12} we expanded the definition of a revision operator to allow for a more detailed description of the context and type of modifications they perform. This flexibility allows us to use these operators to represent patient preferences without any changes to their definition.

Formally, a revision operator $RO^k$ is defined as:

$$RO^k = (\alpha^k, Op^k)$$

where $\alpha^k$ is a logical sentence that describes the interaction applicability of the operator to the theory $\mathcal{D}_{\text{cpg}}$, and $Op^k$ describes the revisions introduced by $RO^k$. In particular, $Op^k$ is a set of $n$ pairs of formulas $(\varphi^k_i, \phi^k_i)$ $(i = 1 \ldots n)$ that define a single operation within the operator. These operations are applied only to $\mathcal{D}_{\text{cpg}}$, so other components of $\mathcal{D}_{\text{comb}}$ are protected from unwanted revisions. For example, $\mathcal{D}_{\text{pi}}$ is never modified thus patient information is never inadvertently changed. The formulas are interpreted as follows (where $\emptyset$ indicates an empty formula):

- $(\varphi^k_i, \emptyset)$ means that $\varphi^k_i$ is removed from any sentence in $\mathcal{D}_{\text{cpg}}$ where it appears,
- $(\emptyset, \phi^k_i)$ means that $\phi^k_i$ is added as a new sentence to $\mathcal{D}_{\text{cpg}}$,
- $(\varphi^k_i, \phi^k_i)$ means that $\varphi^k_i$ is replaced by $\phi^k_i$ in any sentence in $\mathcal{D}_{\text{cpg}}$ where it appears.

Checking the applicability of $RO^k$ to $\mathcal{D}_{\text{comb}}$ translates into the entailment problem $D_{\text{comb}} \models \alpha^k$. We demonstrate in the Applying Preferences section how we leverage this property in our mitigation algorithm.
Figure 1 shows the general architecture of our mitigation framework. It involves two levels of operation -- FOL and regular expressions (RE), i.e. string patterns, often used in find and replace-type operations on text in document retrieval, data analytics and database queries. The core component is the mitigation algorithm, presented later in detail, that operates at the FOL level. The algorithm takes as input a combined theory \( D_{\text{comb}} \), uses revision operators \( RO_k \) to mitigate (identify and address) adverse interactions, and finally constructs a therapeutic scenario \( D_{\text{th}} \) -- a theory that represents a “safe” (free of adverse interactions) course of action. \( D_{\text{th}} \) highlights the clinical actions to be taken (along with the order in which they should be carried out), and includes assumptions made about the patient's state.

**Previous work on Applying Patient Preferences**

Patient preferences and values are considered the third important component of evidence-based medicine\(^1\), the other two being research evidence and clinical experience. Incorporation of patient preferences is especially relevant when dealing with “gray zone” decisions associated with a high level of uncertainty\(^2\). This uncertainty is associated with insufficient evidence\(^3\), evidence indicating the same attractiveness of multiple options or significant differences in preferences across target population. As such these decisions are often referred to as “preference-sensitive”\(^3\). Patient preferences are also relevant when deciding on long-term treatments (for example in case of diabetes, hypertension or osteoporosis)\(^4\). Given the importance of patient preferences, it is now advocated that patients should not only be involved in making decisions when a CPG is being followed, but they should also participate in the development of CPGs\(^5\).

There has been significant research on decision aids that help patients express their preferences (see the next section for more details)\(^1\). Depending on the applied approach to treatment decision making (shared or informed), preferences are considered by the physician and patient or by the patient alone when manually making a specific treatment decision\(^1\). A trial of decision aid tools (implemented as paper charts) for patients considering total knee arthroplasty was described in \(^1\). According to the results of the trial, patients exposed to decision aids demonstrated significantly higher quality decisions (informed choice -- surgery vs. non-surgery -- that matched their values for outcomes of options) than patients that did not use decision aids.

There is also ongoing work on embedding patient preferences in clinical decision support systems. This line of work is exemplified by the MobiGuide system\(^2\). MobiGuide supports remote and CPG-based patient management. The system first learns patient preferences (given in the form of utilities) and applies them when making preference-sensitive decisions in order to establish support for specific options. The system currently contains CPGs for atrial fibrillation\(^3\) and gestational diabetes\(^2\).

**Eliciting Preferences**

Patient preferences represent “the desirability of a health-related outcome, process or treatment choice.” Patients might prefer less invasive tests or procedures over those that cause discomfort, tests that are carried out by nurse practitioners versus those performed by physicians, or specific drugs over others within the same drug class. Such preferences are often represented using “health utilities”\(^6\) that are elicited using various decision aids integrated with
CPGs. These aids typically take the form of charts where patients assign weights or utilities to available options. However, there is no consensus whether this is the best approach that should be adopted in practice.

In our research we use a different approach to elicit patient preferences -- instead of asking patients for direct specification of utilities associated with individual options, we ask them to compare pairs of selected options (for example those for which they have pronounced preferences). These comparisons are then used to derive a preference model in the form of an additive value function that is a sum of marginal value functions associated with specific features of considered options (e.g. their cost, complexity, or side effects). For this purpose we employ the Generalized Regression with Intensities of Preference (GRIP) method that not only constructs value functions, but also allows for detailed analysis of pairwise comparisons and for indicating conflicting or inconsistent answers. This makes GRIP very well suited for interactive use and an example of using GRIP to elicit patient preferences related to pain management therapies is presented in 22.

Value functions constructed by GRIP can be used to evaluate options, also those not included in pairwise comparisons. In our framework we use these evaluations to define desired revisions to CPGs. Specifically, we consider pairs of options for a given action, i.e. a default option and its alternative. If an alternative option receives a better evaluation than the default one, we introduce a revision that replaces the default option with the alternative (see Representing Preferences section below for more information).

Representing Preferences

Up to this point, our mitigation framework only used revision operators to represent secondary medical knowledge (such as drug-drug or drug-disease adverse interactions) that, when relevant to the patient encounter, must be applied to the combined theory used to construct the therapeutic scenario for the patient. Support for preferences introduces the notion of modifications to the combined theory that are desired but not required. Therefore, the mitigation framework must distinguish between these two classes of revisions and apply all applicable interaction-related revisions, while it can skip any of the triggered preference-related revisions that medically invalidate the combined theory.

In this work, we describe revisions as one of two types: those related to preferences and those related to adverse interactions. As such we broaden the definition of an operator to cover all secondary knowledge to be considered when mitigating concurrently applied CPGs. Using this broader definition of a revision operator, we introduce two classes of operators: preference-related revision operators and interaction-related revision operators. The class distinctions are used to define what secondary knowledge must be contained in the combined theory and what secondary knowledge is desired by not required in the combined theory. We note however that the framework uses the same formalism to represent both classes. Therefore, preference-related as well as interaction-related operators can replace, remove, or add clinical actions.

Specific examples of both classes of revision operator are given in the Case Study section. While the preference-related revision operator example and its representation is quite simple, our mitigation framework supports more complex revisions. Preferences for drugs can include dosage adjustment, ordering changes (taking some drug before/after another) and other supported revisions as further described in 7.

Applying Preferences

Providing a formal representation for preferences enables our mitigation framework to apply them while constructing a therapeutic scenario. As integral components, preferences must be vetted medically to ensure no adverse interactions are introduced as a result of their application. Furthermore, formally representing preferences means they can be expressed across a broad range of therapy characteristics. Not only can they refer to higher level concepts such as drugs or medical tests, but preferences can be defined for the type of drug administration (oral, injection, etc.), the number of times the drug is taken per day, if a drug is covered/not covered by insurance, and other similar characteristics. The power of FOL allows our mitigation framework to expand the vocabulary as needed to increase the richness of preferences. In this section we describe how preferences are applied by an extended mitigation algorithm while maintaining medical validity of the proposed therapy.

Our existing mitigation algorithm is a core component of the mitigation framework and it is described in pseudocode in Figure 2. The algorithm takes as input a combined theory $\mathcal{D}_{comb}$, uses interaction-related revision operators $RO^k$ to mitigate (identify and address) adverse interactions, and finally constructs a therapeutic scenario $\mathcal{D}_{th}$. The algorithm iterates over available interaction-related revision operators (line 3). It checks through entailment whether the current operator $RO^k$ is applicable to $\mathcal{D}_{comb}$ (line 5) using a theorem prover. If $RO^k$ is applicable, then the
The mitigation algorithm in Figure 2 employs interaction-related revision operators. When applicable, they must be used so that the therapeutic scenario $D_{th}$ can be constructed. With the introduction of preference-related revision operators, whose application is desired but not strictly required to produce the therapeutic scenario, we extend the mitigation algorithm to treat each type of revision operator differently. The key innovation lies in an extended procedure that acts as a wrapper around the existing mitigate procedure.

Our extended algorithm is more formally shown in Figure 3 as procedure customize_and_mitigate. As input it takes the combined theory $D_{comb}$ and outputs the therapeutic scenario $D_{th}$ that includes changes introduced by all applicable interaction-related revision operators and those preference-related revision operators that do not conflict with the CPGs and with interaction-related revisions. The extended algorithm first orders preference-related revision operators $RO^k$ in decreasing priorities (line 2), as identified by the patient. It then iterates over the ordered operators (line 3), considering more important ones first. For each preference-related $RO^k$, the algorithm checks its applicability through entailment (line 5). If $RO^k$ is applicable, then the algorithm creates a temporary combined theory $D_{comb}'$ by revising $D_{comb}$ (line 7). If $D_{comb}'$ is consistent (i.e. the preferences represented by $RO^k$ do not conflict with the CPGs), then it is passed to the mitigate procedure defined in Figure 2 to apply relevant interaction-related revision operators and to ensure the existence of a therapeutic scenario $D_{th}$. If $D_{th}$ exists, the preference-related revisions are preserved (lines 13-14), otherwise they are discarded due to their conflict with interaction-related revisions. The extended algorithm also invokes the mitigate procedure if no preference-related revision operators were successfully applied, due to no applicable preference operators or applied operators only resulted in conflicts (line 19-20). This case is equivalent to the operations carried out by our existing mitigate algorithm.

Here we note that because the extended mitigation algorithm sequentially applies preference-related revision operators in order of decreasing priority, it is possible that preference-related revisions introduced earlier could “block” preference-related revision operators with lower priorities. Still, the algorithm attempts to consider as many
preference-related revision operators as possible given it does not terminate when a specific operator introduces conflicts, but instead skips over this operator. The framework also assumes that it is better to select the therapeutic scenario $D_{th}$ that contains paths through the CPGs that do not require a preference-related revision to be made. Intuitively, this means the framework elects to return the scenario with as many actions as possible to which the patient has no applicable preference-related revision operators. Our expanded mitigation approach, with respect to preferences, assumes when a patient does not express preference-related revisions for a given clinical action, she fully accepts what is suggested by the CPG.

```
procedure customize_and_mitigate(inout D_comb, out D_th) begin
    preferences_applied := false;
    order preference-related RO's according to their decreasing priorities;
    foreach preference-related RO do
        begin
            if $D_{comb} = a^k$ then
                begin
                    $D_{comb}' :=$ revise $D_{comb}$ using $Qp^k$;
                    if $D_{comb}'$ is consistent then
                        begin
                            mitigate($D_{comb}'$, $D_{th}$);
                            if $D_{th} <>$ null then
                                begin
                                    preferences_applied := true;
                                    $D_{comb} := D_{comb}'$
                                end
                            end
                        end
                    end
                end
            end
        end
    if not preferences_applied then
        mitigate($D_{comb}$, $D_{th}$);
end
```

**Figure 3.** Extended mitigation algorithm.

**Case Study: Management of a Patient with Stable Cardiac Artery Disease Who Suffers a Sudden Onset of Deep Vein Thrombosis**

Stable coronary artery disease (SCAD) encompasses several patient populations at different stages and with different types of coronary disease, excluding acute coronary syndromes. SCAD has a number of manifestations, including some form of chest discomfort (the most common one) that is induced by exercise, emotion, or stress. People with hypertension, hypercholesterolemia, diabetes, obesity, and those who smoke are at a higher risk of developing SCAD\(^23\). Treatment of a typical SCAD patient begins with short acting nitrates combined with beta blockers or calcium channel blockers as well as lifestyle management to control for cardiovascular risk factors. If SCAD symptoms persist, a second line of treatment is initiated that includes long lasting nitrates combined with potassium channel activators (e.g. nicorandil), statins, and aspirin.

In the clinical scenarios described below we assume a SCAD patient presenting with symptoms of deep vein thrombosis (DVT) related to either DVT itself or a pulmonary embolism (PE) -- we refer to this presentation as DVT/PE. A patient diagnosed with DVT/PE is managed according to the standard CPG\(^24\) and a simplified version of this guideline is presented in Figure 4 as an actionable graph\(^10\) (we omit the figure showing the SCAD CPG due to space limitations but it is available from the authors upon request). The management of DVT/PE involves aggressive in-patient therapy especially for those patients who exhibit haemodynamic instability or are diagnosed with renal failure. Such a therapy is followed by out-patient management. The development of novel oral anticoagulants (NOACs) opened up new treatment options beyond management with low molecular weight heparin (LMWH) and warfarin. It is in the context of NOACs that patient preferences start playing an important role in selecting the most suitable and clinically sound treatment. In the case of DVT/PE out-patient management, a patient can consider one of the following therapeutic options:
LMWH and warfarin; this is the default treatment (and presented in Figure 4) that requires regular blood work in order to assess the effect of oral anticoagulation with warfarin. This assessment is done using the international normalized ratio (INR).

LMWH for 5 days followed by NOAC (dabigatran) twice daily; some patients may prefer to start with a parenteral agent (e.g., LMWH injection) as it is perceived as a more effective immediate treatment. No blood work is required for this option.

NOAC alone; possibilities include either rivaroxaban that is taken once a day or apixaban that is taken twice a day and has a slightly lower risk of bleeding than rivaroxaban. Treatment with NOACs does not require blood work.

Figure 4. DVT/PE CPG represented as an actionable graph (DVT = deep vein thrombosis, PE = pulmonary embolism, UFH = unfractioned heparin, LMWH = low molecular weight heparin, IVCF = inferior vena cava filter)

Patient preferences governing the selection of a given treatment option may include such factors as perceived effectiveness (injection vs. oral administration), ease of use (injection vs. oral vs. number of daily administrations), convenience (no need to monitor effects of anticoagulation, no need to buy and store syringes), or insurance coverage of medication (government, private, or out of pocket expense). In the subsequent sections we illustrate how patient preferences are modeled as part of our mitigation framework and how they impact the development of a therapeutic scenario. We start with a simple illustration of mitigation that does not include patient preferences and finish with one that does.

Clinical Scenario 1: Mitigating Adverse Interactions

In the first scenario we assume an elderly male patient is treated for recurring SCAD and has not expressed any preferences for their DVT/PE treatment. According to the SCAD CPG, he is placed on long lasting nitrates, potassium channel activators, statins, and aspirin. This patient arrives at the Emergency Department (ED) complaining of swelling and tenderness in his left leg. An ultrasound and D-dimer test confirm DVT. Further
investigations rule out haemodynamic instability and renal failure. The patient is immediately started on anticoagulation therapy (unfractioned heparin (UFH) and warfarin).

However, a patient should not be put on two different anticoagulation treatments (in this case aspirin or clopidogrel for SCAD and any type of anticoagulation for DVT/PE). Supporting this scenario requires the codification of secondary knowledge preventing the use of two anticoagulation therapies. As stated in medical literature for patients treated for DVT/PE, the administration of aspirin and clopidogrel must be stopped when the patient is put on another anticoagulation therapy. The interaction-related revision operator $RO^1$ represents the secondary knowledge that can be used to mitigate this drug-drug interaction. Note we use a simplified representation that does not use variable names but only presents their labels:

$$RO^1 = \langle \alpha^1, Op^1 \rangle,$$

$$\alpha^1 = \text{diagnosed(DVT\_PE)} \land \text{execute(anticoag)} \land \text{execute(aspirin)},$$

$$Op^1 = \langle\text{execute(aspirin)} \land \text{execute(clopidogrel)}, \emptyset \rangle.$$  

For this scenario we have the following subset of sentences describing the patient’s state $D_{pi}$.

$$D_{pi} = \text{diagnosed(DVT\_PE)}, \text{diagnosed(SCAD)}, \text{execute(aspirin)}.$$  

The following sentences are part of the theories for SCAD and DVT/PE, including a sentence describing the applied anticoagulation therapy that uses warfarin and unfractioned heparin (UFH).  

$$D_{SCAD}^{DVT\_PE} \ni \text{execute(clopidogrel)}$$

$$D_{cplg}^{DVT\_PE} \ni \text{execute(anticoag)} \implies \text{execute(warfarin)} \land \text{execute(UFH)}$$  

To check for the applicability of $RO^1$ we formulate the entailment problem $D_{comb} \models \alpha^1$. Because $\alpha^1$ is entailed by $D_{comb}$, we immediately establish that $RO^1$ is applicable to $D_{comb}$ and we consequently apply $Op^1$ to remove aspirin and clopidogrel from $D_{cplg}^{SCAD}$. This resolves the drug-drug interaction and the returned therapeutic scenario $D_{Th}$ includes the administration of long lasting nitrates, potassium channel activators, statins, but not aspirin or clopidogrel for the subset of the model for $D_{Th}$ related to treating the patient’s SCAD. Had the entailment problem $D_{comb} \models \alpha^1$ failed, we would be unable to revise $D_{comb}$ and find a model for it. Similarly, because no preferences were expressed by the patient, only interaction-related revision operators are checked and applied. We omit the full theory describing this patient encounter due to space limitations and refer the reader to previous work for more detailed scenarios \cite{3,4,12}.

**Clinical Scenario 2: Applying Patient Preferences**

Considering the same patient as in the first scenario, now the patient has expressed a preference for a simple treatment of DVT/PE, defined as the fewest drug administrations in a day as possible, oral route preferred over injection, and the administration of drugs covered by insurance (the patient has supplementary private health insurance). Due to the blood work required before administering warfarin, the patient prefers to take rivaroxaban. Taking this drug allows the patient to start with pills only (no parenteral agent), does not require blood work, the pill is taken only once a day, and is covered by the government (e.g., for seniors) and private insurance. This is a common patient preference as indicated by the medical expert (Dr. Carrier) on our team. The preference-related revision operator $RO^2$ represents, in a structured form, the above expressed patient preference that was elicited prior to prescribing a treatment. Note again we use a simplified representation to improve readability:

$$RO^2 = \langle \alpha^2, Op^2 \rangle,$$

$$\alpha^2 = \text{diagnosed(DVT\_PE)} \land \text{execute(anticoag)},$$

$$Op^2 = \langle\text{execute(warfarin)} \land \text{execute(UFH)}, \text{execute(rivaroxaban)} \rangle.$$  

To check for the applicability of the preference-related revision operator $RO^2$ we first formulate the entailment problem $D_{comb} \models \alpha^2$. Because $\alpha^2$ is entailed by $D_{comb}$, we immediately establish that $RO^2$ is applicable to $D_{comb}$ and we consequently apply $Op^2$ to replace the anticoagulation therapy that uses warfarin and UFH with one that uses rivaroxaban in $D_{cplg}^{DVT\_PE}$. Note that this results in the sentence $\text{execute(anticoag)} \implies \text{execute(rivaroxaban)}$ as part of $D_{comb}'$. Applying the $\text{mitigate}$ procedure to $D_{comb}'$ results in a revision made according to $RO^1$ since $\alpha^1$ is still entailed by $D_{comb}'$. As such, the returned therapeutic scenario $D_{Th}$ both resolves the drug-drug interaction
described in RO1 and applies the patient’s preference for a simpler treatment of his DVT/PE condition as represented by RO2. We again omit the full theory describing this patient encounter due to space limitations.

Discussion

In this paper we describe extensions to our FOL-based mitigation framework to add support for patient preferences when mitigating multiple concurrently applied CPGs to a multimorbid patient. The flexibility provided by FOL allowed us to easily expand the definition of an operator to include two subclasses that use the same representation formalism as before. We leverage these classes of operators in an extended mitigation algorithm to either apply all revisions represented by the set of interaction-related revision operators or as many as possible from the set of preference-related revision operators. We also describe how preferences are formally elicited, further demonstrating our incorporation of preferences as first-class citizens in the mitigation process.

In supporting preferences, we demonstrate our mitigation framework’s ability to represent different types of secondary knowledge. The current version of the framework supports preferences that replace a medical action(s) with one or more different ones. As future work we will support the full range of operations for preference-related revision operators. Additionally we are further working on the mitigation framework’s theoretical foundations that will make it possible to expand the notion of mitigation to include preferences, with the goal of developing an interactive clinical decision support system (CDSS) to be used at the point of care.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada, Telfer School of Management Research Fund. The authors acknowledge support from Adventium Labs for the first author.

References


Design of a Community-Engaged Health Informatics Platform with an Architecture of Participation

Mari Millery, PhD¹, Wilson Ramos, MSEE¹, Chueh Lien¹, Alejandra N. Aguirre, MPH², Rita Kukafka, DrPH, MA, FACMI¹

¹Columbia University Department of Biomedical Informatics, New York, NY; ²Irving Institute for Clinical and Translational Research, Columbia University Medical Center, New York, NY

Abstract

Community-engaged health informatics (CEHI) applies information technology and participatory approaches to improve the health of communities. Our objective was to translate the concept of CEHI into a usable and replicable informatics platform that will facilitate community-engaged practice and research. The setting is a diverse urban neighborhood in New York City. The methods included community asset mapping, stakeholder interviews, logic modeling, analysis of affordances in open-source tools, elicitation of use cases and requirements, and a survey of early adopters. Based on synthesis of data collected, GetHealthyHeights.org (GHH) was developed using open-source LAMP stack and Drupal content management software. Drupal’s organic groups module was used for novel participatory functionality, along with detailed user roles and permissions. Future work includes evaluation of GHH and its impact on agency and service networks. We plan to expand GHH with additional functionality to further support CEHI by combining informatics solutions with community engagement to improve health.

Introduction

In biomedical informatics, the level of community remains a fairly unexplored application domain. While public health informatics generally addresses population health, community-level engagement is rarely investigated from an informatics perspective. Community-based participatory approaches have rarely been applied in public health informatics. Meanwhile, there is increasing recognition that how people live in their communities is a major determinant of health. Recent healthcare policies in the U.S., such as patient-centered medical homes and population-based accountable care initiatives, are consistent with the idea of situating patients and health resources in communities. Academic medical centers, such as the Columbia University Medical Center in New York City where the present study is located, are moving towards the population health model required by the Affordable Care Act by recognizing the value of partnering with community stakeholders and leveraging a broad spectrum of community assets.

Goodman and colleagues provide a useful definition of community health as “a multi-sector and multi-disciplinary collaborative enterprise that uses public health science, evidence-based strategies, and other approaches to engage and work with communities, in a culturally appropriate manner, to optimize the health and quality of life of all persons who live, work, or are otherwise active in a defined community or communities.” The potential power of community-level health interventions is increasingly recognized. The characteristics of the ecological systems paradigm for community interventions encompass a focus on community capacity and issues identified through community engagement; ecological and systemic perspectives; empowerment of the community; and the permeating role of culture.

Community-engaged health informatics (CEHI) applies information technology and participatory approaches to improve the health of communities. CEHI facilitates community participation in research and dissemination of evidence. The emerging agenda of CEHI incorporates concepts and methodologies from biomedical informatics, community-based public health approaches, and other fields such as community informatics. Community informatics is the application of information and communication technology to enable and empower community processes. CEHI applies tenets of community informatics in the domain of health. Community-based participatory research (CBPR) is a public health research approach defined as a partnership approach to research that equitably involves community members, organizational representatives, and researchers in all aspects of the research process and in which all partners contribute expertise and share decision making and ownership. Interestingly, CBPR shares many aspects with participatory approaches to information technology design that actively involve end-users in the design process.
Community health interventions can be more effective and sustainable if they build on existing community capacity. CBPR emphasizes community asset mapping. Community-based organizations (CBOs) and service providers are essential assets for building community health interventions. They are important stakeholders in health systems because they provide a wide spectrum of programs and services to members of their community, link with other health and social services to help provide care, and advocate for broader system-level supports. Emerging literature in the area of knowledge translation is also beginning to highlight the potential of CBOs in translating evidence into the community. Many barriers hinder evidence use and evidence-based practices of CBOs. Research has consistently identified that CBOs struggle with: access to evidence; time to process evidence; skills to review, summarize, and synthesize evidence; research terminology; and local applicability and acceptability of evidence. Strategies are needed to address these barriers in order to realize the potential of CBOs as important agents of improving population health.

CEHI takes an ecological systems approach to community health. It extends the notion of Learning Health System to the community-level, and investigates the ecology of community health information in a “cyber-social ecosystem.” The present community intervention seeks to integrate a participatory online environment in the setting of a real geographic community environment. The rationale of the intervention is supported by evidence that online participation and public/civic participation are found to mutually reinforce each other. While many online health communities have broken down geographic barriers by connecting people with similar health interests, it will also be important to build online health communities that have great potential to support health promotion in real geographically defined communities.

The objective of our research was to translate the concept of CEHI into a usable and replicable model informatics platform that will facilitate community-engaged practice and research. The goal of the research is to generate generalizable knowledge and tools that apply to communities outside of Washington Heights/Inwood (WAHI). GHH in WAHI is used as a “laboratory” environment to develop the concept of the CEHI platform. Plans are currently underway to bring CEHI platforms to other communities. This paper describes the methods and results of the GHH CEHI platform design process.

Methods

Setting

The setting of this study, the Washington Heights/Inwood (WAHI) neighborhood, is located in upper Manhattan within New York City (NYC), north of Harlem and directly south and west of the Bronx. It is a densely populated urban area with approximately 280,000 residents. A large proportion of the community residents are Hispanic (71%) and nearly 90% belong to a racial/ethnic minority group. African Americans represent 14% of the population. Most residents are foreign-born, mainly from Latin America, with the vast majority from the Dominican Republic (71%), followed by Ecuador (4%). Less than 50% are proficient in English. The median household income in 2007 was $35,456, and 27% lived below the 200% federal poverty level in 2007. A large percentage (44%) of community residents (persons over the age of 25) did not graduate from high school, and the unemployment rate is over 12%. The most common sources of employment are services and sales industries. Health concerns in the community are significant compared to NYC as a whole. One third (32%) of WAHI residents rate their health as fair or poor, compared to 21% citywide. Residents are less likely to have a regular doctor, more likely to seek care from the emergency department, and more likely to be uninsured. The community rates poorly regarding risk factors for heart disease, with higher rates of high blood pressure, high cholesterol, obesity and diabetes than the overall NYC population, combined with less recreational exercise.

Design Process

Design process is iterative, unique in each case, and may incorporate a variety of practices from different fields. The GHH design process involved a core team at the Department of Biomedical Informatics at Columbia University, including all authors of this paper, and ongoing consultation with community-based organizations and other stakeholders. The idea for GHH originally emerged from lessons learned from another community health portal, GetHealthyHarlem.org, which was previously developed under Dr. Kukafka’s leadership. Whereas GetHealthyHarlem.org was designed primarily for dissemination of local health information to community members, GHH was designed for multiple levels of the community--from community residents to CBOs and other
community stakeholders—to share health related content and connect with each other. The GHH team sought to develop a platform that would facilitate community engagement in health practice and research. CBOs with a health-related mission were recruited as initial primary users. A full range of other stakeholders and related community sectors may also participate in GHH, including, for example, schools, law enforcement, urban planning, and religious organizations.

Following is a list of key steps in the CEHI platform design process for GHH. It is important to note that the steps often deviated from the numeric order shown here, with iterative loops repeating prior steps and with overlapping steps conducted simultaneously.

1. Preliminary community asset mapping
   The design of GHH is supported by Columbia University’s Irving Institute for Clinical and Translational Research, which is the Clinical and Translational Science Award (CTSA) at Columbia University. The Community Engagement Core Resource (CECR) of the Irving Institute conducts extensive outreach and community asset mapping in WAHI, which is the neighborhood where Columbia University Medical Center (CUMC) and the Irving Institute are located. The relationships CECR has established with community-based organizations and other stakeholders are the foundation for the community-based participatory approach of our design process. CECR is also responsible for ongoing site management and promotion, and engagement of organizational users, including coordination of the GHH Steering Committee.

2. Logic modeling
   As is common practice in public health program planning, the team constructed a logic model to represent inputs, activities, outputs, and outcomes envisioned for GHH.

3. Guiding principles for CEHI platform
   The team reached consensus on several guiding principles that helped align the work with a common understanding of CEHI and architecture of participation.

4. Key informant interviews
   Six semi-structured interviews were conducted with key stakeholders, including community leaders, executives of community-based organizations, and leaders of CUMC groups that engage in community-based participatory research. The interviews covered a broad range of questions relevant to GHH planning, but in this paper we present key themes related to functional needs expressed by the interview participants. Several team members engaged in the collaborative qualitative coding process to derive these themes.

5. Identification of open-source tools
   One of the guiding principles (see Step 3) was to use open-source tools when possible. The team identified suitable tools for the GHH CEHI platform and decided to use Drupal as the primary development environment.

6. Exploration of affordances in Drupal
   The team reviewed Drupal documentation, engaged a Drupal developer consultant, and explored design features in publicly available health and non-health websites designed using Drupal.

7. Use cases and requirements
   Step 7 occurred simultaneously with Step 6. The team analyzed use cases, partly informed by the key informant interviews, and articulated most important requirements for a CEHI platform, given affordances available in Drupal.

8. Iterative development process
   Wireframes and prototypes were developed iteratively, working closely with a Drupal expert consultant.

9. Definitions of roles and permissions
   Once the team decided to use Drupal Organic Groups module for the essential participatory functionality, a thorough process was undertaken to configure user roles and permissions.

10. Quality testing
    Student interns and other team members conducted a systematic quality assurance process of all features of GHH over several months.

11. Soft launch with early adopters
    An initial group of 10 organizations were invited to participate in the soft launch of GHH.
12. Survey of early adopters
   During a session that introduces GHH and reviews key functionality, early adopters (n=18) completed a structured survey asking about user needs and reactions to GHH functionality.

13. Plan for full launch
   A detailed plan was developed for full community-wide launch of GHH, including several strategies of outreach and marketing.

Results
The mission of the GHH CEHI platform was collectively defined by the GHH team and early adopters as “an online community that engages people in Washington Heights-Inwood to connect, discover, and share resources to get healthy”. Figure 1 shows one part of the GHH logic model, representing the sequence of outcomes leading from use of GHH to intermediate outcomes and then to improved community health as the long-term outcome. A degree of shared understanding about the mission, intermediate outcomes, and long-term outcomes was achieved early on in the design process.

![Figure 1. Logic of GHH outcomes.](image)

The team had extensive discussions to clarify basic principles to guide the design of a CEHI platform. One of them was to use a CBPR approach and to incorporate community stakeholder input and engagement throughout the process to a maximum extent possible. Another principle was to utilize open-source technologies when possible, given their multiple advantages over proprietary tools, such as customizability, flexibility, interoperability, and support options. An essential principle underlying all functional requirements was to strive for opportunities and strategies to increase active user participation. To that end, the team reviewed dozens of publicly available websites, with a special emphasis on health-themed sites, and listed ideas for functions that facilitate contributions from and engagement by users.

Semi-structured key-informant interviews were used to further elicit input from important stakeholders. The six key informants included community leaders, executives of community-based organizations, and leaders of community-engaged research and service providers from the medical center. Table 1 shows key themes that reflect functional needs the informants expressed for accomplishing their work with the WAHI community. Six themes are listed, with sub-themes under each. There was a high degree of agreement across the different types of key informants. For example, both community-based and medical center-based informants emphasized the first five main themes listed in Table 1. Although medical center stakeholders expressed more needs around research production and use, research related needs were also discussed by community-based participants. The interview data provided the team an understanding of higher-level functional needs among potential users.
Table 1. Qualitative interview coding themes for functional needs among key informants.

<table>
<thead>
<tr>
<th>Main Theme</th>
<th>Sub-Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connections with organizations and people</td>
<td>Networking</td>
</tr>
<tr>
<td></td>
<td>Finding collaboration partners</td>
</tr>
<tr>
<td></td>
<td>Knowing who is doing what/working on what issues</td>
</tr>
<tr>
<td></td>
<td>Ongoing conversation and exchange</td>
</tr>
<tr>
<td>Information on community needs</td>
<td>Awareness of problems/issues affecting community</td>
</tr>
<tr>
<td></td>
<td>Access to community-specific data</td>
</tr>
<tr>
<td>Information on community resources and services</td>
<td>Up-to-date, comprehensive, and trusted list for referrals</td>
</tr>
<tr>
<td></td>
<td>Having a place to go to find a current health related resource</td>
</tr>
<tr>
<td>Information on health</td>
<td>Useful health education materials</td>
</tr>
<tr>
<td></td>
<td>Current expert information on health issues</td>
</tr>
<tr>
<td></td>
<td>Desire to share one’s expertise</td>
</tr>
<tr>
<td>Communication with constituents</td>
<td>Gaps in existing communication methods</td>
</tr>
<tr>
<td></td>
<td>Marketing in community</td>
</tr>
<tr>
<td>Research production and use</td>
<td>Recruitment</td>
</tr>
<tr>
<td></td>
<td>Dissemination</td>
</tr>
<tr>
<td></td>
<td>CBPR participation</td>
</tr>
</tbody>
</table>

The next step in the process described in the method section (Step 5) involved identification of suitable open-source technologies. Based on discussions with experts and examination of functionality, Drupal open-source content management system was selected as the primary tool for building the GHH CEHI platform due to its rich functionalities, support for multilingual sites, support for multiple content types, advanced user management that allows for multiple site stakeholders, stability, and scalability. Open source LAMP stack (Linux operating system, Apache HTTP server, MySQL relational database management system, and PHP programming language) was used as software components of the underlying platform. An iterative process (Steps 6-8) was undertaken with a Drupal developer consultant to explore Drupal features, refine use cases and requirements, and design several iterations of wireframes.

In response to the findings from use cases and key-informant interviews, the team selected an essential-set of functions to engender community engagement. Selected functions included a community calendar, a local service directory, posting of multiple types of content (e.g., articles, videos, and links), the ability to comment and rate content, integration of social media for content sharing, use of Google Translate (especially for Spanish translation of content), and the creation of social networking groups. For example, the local service directory functionality relates to the interview theme “Information on community resources and services” (Table 1), and the interview theme of “Information on health” motivated the functionality for posting content. The social network group functionality was deemed an essential design point of GHH. It entailed forming subgroups within the general community based on a localized interests or purposes. Groups could be public or private in nature, with the ability to determine their own membership, host relevant content and provide feedback or communicate with its members. A core role of GHH as a CEHI platform revolved around providing this group functionality to the WAHI community. This functionality meets requirements derived from two particular themes listed in Table 1: “Connections with organizations and people” and “Communication with constituents”.

To enable this function, GHH chose to deploy Drupal 7.1 Organic Groups module. The ability to dynamically form groups is core to the social network group functionality of GHH. Groups within GHH are defined by users and have access to communication forums, email, blogging and notification services, content-sharing services, event calendars and links to social media sites. Ownership and management of the group is kept at the community-level through the pre-defined role of a group owner. Specific group features were defined at both the content creation and the content consumption level. By design, Organic Groups also links to Drupal administrative and reporting services, thus allowing GHH to monitor and provide access and security to groups from an overall site perspective. The full vision of GHH is to support local groups within the WAHI community. These groups must be able to dynamically form, grow, shrink and even disappear according to community preferences. Through Drupal Organic Groups, GHH provides this ability online and fulfills a crucial tenet of community-based participatory research.

Given the required type of participatory architecture of GHH, definition of user roles and permissions also became a complex and important design task for the team. Over several weeks, the team constructed a set of user role and permission settings for the GHH site as a whole (Table 2), and sets of permission settings for public and private groups by type of group member status. The public group settings are shown in Table 3.
Numerous issues were discovered during the quality-testing phase. The user roles and permissions in particular required extensive testing and re-testing to ensure functioning of the exact specifications listed in Tables 2 and 3. Other issues involved appropriate requirements for file sizes and formats of content that users can post. Issues of security, including CAPTCHA verification, also required attention. Overall, our experience demonstrated the value of extensive testing and gradual launch process.

**Table 2.** Drupal permissions settings for four user roles on GHH.

<table>
<thead>
<tr>
<th>GHH Member Status</th>
<th>Anonymous User</th>
<th>Unvalidated User</th>
<th>Authenticated User</th>
<th>Group Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>A non-registered user who visits GHH.</td>
<td>Has registered for GHH but has yet to authenticate their account by using the link in their email confirmation.</td>
<td>A registered user that has authenticated their account.</td>
<td>Responsible for managing group. Role can be assigned to more than one member.</td>
</tr>
</tbody>
</table>

**Usage Permissions**

<table>
<thead>
<tr>
<th>Permission</th>
<th>Anonymous User</th>
<th>Unvalidated User</th>
<th>Authenticated User</th>
<th>Group Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can Cancel Own Account</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Search on Site</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View Public Content Across Site</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View Media on the Site</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View Comments</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Post Comments</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Request Group Membership</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Can Request to Create a Partner Organization or Interest Group</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Post Content</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Comment on Group Content</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Post Under Group Name</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View User Profile</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Contact Users Via Contact Form</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Attach Files to Contact Forms</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View and Use Users Email Address</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View Group Members List</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View Group Members’ Profiles</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Manage Group Members’ Status</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Table 3.** Drupal Organic Groups permissions settings by group member status for public GHH groups.

<table>
<thead>
<tr>
<th>Group Member Status (only for registered GHH users)</th>
<th>Non-member</th>
<th>Pending Member</th>
<th>Active Member</th>
<th>Blocked Member</th>
<th>Group Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>User has not requested group membership. Group Manager needs to change status to Active.</td>
<td>Group Manager has approved request for membership by assigning Active status to member.</td>
<td>Group Manager has blocked member from group.</td>
<td>Responsible for managing group. Role can be assigned to more than one member.</td>
<td></td>
</tr>
</tbody>
</table>

**Permission**

<table>
<thead>
<tr>
<th>Permission</th>
<th>Non-member</th>
<th>Pending Member</th>
<th>Active Member</th>
<th>Blocked Member</th>
<th>Group Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can Request Group Membership</td>
<td>Yes</td>
<td>NA</td>
<td>NA</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>Group Listed on Dashboard</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View Group Content</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Headings</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Comment on Group Content</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Post Under Group Name</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Edit Group Content Posted by Him/Her-Self</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Edit Group Content Posted by Other Group Members</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Listed on Group Membership List</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View Group Members List</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can View Group Members’ Profiles</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Contact Individual Group Members</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Access Group Members List-Serve Function</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Manually Add Group Members</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Can Manage Group Members’ Status</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
To provide a feel for the final design product at this stage, Figure 2 shows a screen-shot with some of the participatory functions. (The full website is accessible at http://www.gethealthyheights.org.) On the top left of Figure 2 is the list of content categories used during the soft launch. This list is currently undergoing revision, based on feedback from early adopters. On the bottom left is community event calendar, which has received extremely positive early user feedback. The right hand side shows the area where users can post different types of content (articles, videos, services, links, and events). We will continue an iterative process of gathering further user feedback and refining the functionality. For example, focus groups are planned to obtain feedback from the community member perspective.

![Figure 2. Screen-shot from GHH, illustrating participatory functionality.](image)

Early adopters (n=18) were asked to participate in a brief survey during GHH orientation sessions. One of the survey questions asked users to indicate how likely they think they are to use specific features of the site. The data are shown in Figure 3. The team is currently using this information to further refine functionality. For example, early adopters indicated an interest in linking GHH with social media with 90% responding that they are very likely or somewhat likely to promote GHH on social media (Twitter, Facebook, etc.). As a result, the team is designing more robust features for social media integration. On the other hand, fewer people indicated likelihood to become regular content contributors or post articles, prompting the GHH team to design a strategy for systematically inviting contributions. During the early use period, we discovered that CBO users preferred to create user names on behalf of their organizations, rather than as individuals. This is an example of users appropriating technologies their own ways, sometimes contrary to the designer’s intentions. We decided to allow organizational user names on GHH.
The CEHI platform configurations are presented here as the primary outcome of the research. GHH was launched to an initial set of users from CBOs in April 2015. As of early July 2015, there has been robust use by the CBOs as well as others who have discovered the site. In the first three months, there were 57 registered users, 26 partner organizations, and 5 interest groups on GHH. Table 4 shows further early use statistics of GHH, obtained through Google Analytics. We plan to present updated use data at the AMIA meeting in November 2015, including use patterns of specific functionalities. During the initial months, a steering committee was established as the governing body of GHH. Twelve CBOs are actively participating as part of the steering committee.

Table 4. Early use of GHH.

<table>
<thead>
<tr>
<th></th>
<th>April 2015</th>
<th>May 2015</th>
<th>June 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sessions</td>
<td>777</td>
<td>921</td>
<td>1,198</td>
</tr>
<tr>
<td>Users</td>
<td>473</td>
<td>764</td>
<td>948</td>
</tr>
<tr>
<td>Page Views</td>
<td>3,962</td>
<td>3,209</td>
<td>4,338</td>
</tr>
</tbody>
</table>

Figure 3. Self-reported anticipated likelihood of GHH use scenarios among early adopters.
Discussion

We have described the methods and results of a design process for the GHH CEHI platform with an architecture of participation. Public health methods, such as logic modeling and community asset mapping, were integrated with information technology design methods, including analysis of functional requirements and iterative development of features to meet user needs. Both qualitative interview data and quantitative survey data were utilized as part of the design process. Throughout the process we show how user needs and requirements are translated into platform configurations and features, resulting in a design product that follows the principles of CEHI and is tailored for the local community. In the GHH CEHI platform, configurations for user roles and permissions became central to achieving the desired participatory functionality with the Drupal Organic Groups module. The general model presented here can be applied as a CEHI intervention in any community but needs to be customized for local needs.

CEHI combines the potential of community-engaged public health interventions with the potential of informatics tools. There is little prior work in this intersection of fields, but an increasing demand for innovative community-level interventions has recently emerged in the context of the Affordable Care Act and healthcare reform. The present study addresses the need for community-level interventions and leverages the power of online tools to facilitate community health in the context of geographically defined high-need communities. Other pioneering community health websites have started emerging in the U.S. and other countries. Examples include websites for Sonoma County, CA [http://www.healthysonoma.org], and London, UK [https://www.myhealth.london.nhs.uk], but GHH is distinguished by its degree of participatory architecture and community stakeholder engagement.

While our aim is to develop a platform that can be adapted for a variety of community settings, one limitation of our work is that the characteristics of the GHH CEHI platform may not be suitable for all communities. For example, it is possible that urban settings with relative density of CBOs and health services may benefit from the functionalities in this type of CEHI platform more than rural communities. On the other hand, rural communities may utilize the CEHI platform in a way that is different, requiring modification to the platform’s functionalities. Generalizability of GHH as a CEHI platform will need to be verified in a variety of communities. Another limitation of our work is that we do not have design, programming, and technology infrastructure resources comparable to commercially supported websites.

Future work includes evaluation of GHH use patterns and the impact of GHH on health-related agency and service networks in the WAHI community. We plan to measure community network connectivity outcomes using system-level network analysis techniques. We are currently collecting baseline survey data on the social network connections among agencies in WAHI. The evaluation plan involves measurement of the network connectivity over time and correlating use of GHH with network connectivity outcomes. In order to demonstrate impact on specific community health outcomes, we plan to identify health conditions and specific projects that utilize the GHH platform. Some of the projects under discussion include focus on obesity, breast cancer, and healthy aging. We also plan to develop informatics approaches to further community asset mapping, with the goal of building a robust and updatable community service directory as part of GHH. GHH is intended to be a CEHI platform that allows further exploration of novel functionalities that promote community engagement and community health. Additional desired functionalities include centralized access to data on WAHI and exploration of opportunities to exchange data among community agencies. We will identify best practices for managing and evolving a CEHI platform of this size and scope. A well-structured development and release methodology will be documented based on the GHH team’s experience. After a number of release cycles, the information will be disseminated to the CEHI community to assist in the development of other CEHI sites.

Conclusion

GHH was developed as a model platform for CEHI. This work begins to expand biomedical informatics into the community-level, which is where many important determinants of human health occur. CEHI approaches, such as GHH, are needed to facilitate engagement of community-based assets and resources to promote community health. A CEHI perspective expands the vision of the Learning Health System beyond medical care settings, to fully include community stakeholders in the health data ecosystem.
References

Knowledge Extraction from MEDLINE by Combining Clustering with Natural Language Processing

Jose A. Miñarro-Giménez, PhD, Markus Kreuzthaler, MSc, Stefan Schulz, MD
Institute of Medical Informatics, Statistics, and Documentation,
Medical University of Graz, Austria

Abstract

The identification of relevant predicates between co-occurring concepts in scientific literature databases like MEDLINE is crucial for using these sources for knowledge extraction, in order to obtain meaningful biomedical predications as subject-predicate-object triples. We consider the manually assigned MeSH indexing terms (main headings and subheadings) in MEDLINE records as a rich resource for extracting a broad range of domain knowledge. In this paper, we explore the combination of a clustering method for co-occurring concepts based on their related MeSH subheadings in MEDLINE with the use of SemRep, a natural language processing engine, which extracts predications from free text documents. As a result, we generated sets of clusters of co-occurring concepts and identified the most significant predicates for each cluster. The association of such predicates with the co-occurrences of the resulting clusters produces the list of predications, which were checked for relevance.

Introduction

The state of the art of a scientific discipline and the progress of investigations about a particular topic are described by their set of scientific publications. The identification of new knowledge, which in the past exclusively relied on human effort, has become increasingly difficult due to the accelerating growth in academic literature. Consequently, scholars rely more and more on machine filtering and preprocessing of scientific articles.

The MEDLINE database, with currently about 22 million bibliographic records, is the most important source of biomedical literature. Each record has been semantically annotated by experts of the U.S. National Library of Medicine using the MeSH thesaurus. These metadata are not only useful for document retrieval, but also constitute valuable assets for information and knowledge extraction. They include sets of MeSH terms, which can be further qualified by one or more MeSH subheadings (Table 1), which specify the semantic context of the MeSH term, e.g. Anti-Inflammatory Agents / Therapeutic use, or Nephrotic Syndrome / Drug therapy. As much as this is useful for targeted document retrieval, it falls short of typical knowledge representation use cases, which demand predications like \(<\text{Anti-Inflammatory Agents} \square \text{Treats} \square \text{Nephrotic Syndrome}>\).

The file MRCOC within the UMLS Metathesaurus contains all annotations of co-occurring MeSH terms and subheading qualifiers for each MEDLINE record (aka citation). In the past, the content of MRCOC was used for diverse objectives, such as knowledge extraction, the identification of associative relations, semantic relation discovery, mining of symbolic and statistical gene-disease relationships, and text mining in general.

Another UMLS component is the Semantic Network (UMLS SN), an informal upper-level ontology, which provides 133 generic categories, so-called semantic types, linked by 54 directional relationships. All Metathesaurus concepts are assigned to one or more SN semantic types. Table 2 shows how the semantic types Disease or Syndrome, Organism Function and Pharmacologic Substance are linked in UMLS SN, which, in addition, defines directional relationships such as \(<\text{Disease or Syndrome} \rightarrow \text{Manifestation_of} \rightarrow \text{Organism Function}>\) or \(<\text{Pharmacologic Substance} \rightarrow \text{Disrupts} \rightarrow \text{Organism Function}>\). They are pattern for typical predications, with their subject and object positions to be refined by UMLS Metathesaurus concepts.

The identification of relevant semantic relations is crucial for the generation of predications. Natural language processing (NLP) is the method of choice for extracting such predications from textual sources. One example is SemRep, a system that recovers predications from biomedical text using syntactic analysis and structured domain knowledge from UMLS. However, ambiguity and complexity of biomedical language hinder the accurate extraction of biomedical facts. Capitalizing on the availability of semantic explicit MeSH annotations, we will investigate how this resource can be used to extract factual statements from MEDLINE, despite the lack of relational predications in its metadata annotations.
Table 1. Excerpt of the list of MeSH terms and their related MeSH Subheadings in the sample PubMed record “Childhood nephrotic syndrome – Current and future therapies” (id = 22688744).

<table>
<thead>
<tr>
<th>MeSH term and identifier</th>
<th>MeSH Subheadings and abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nephrotic Syndrome (D009404)</td>
<td>Therapy (TH), Metabolism (ME)</td>
</tr>
<tr>
<td>Anti-inflammatory Agents (D000893)</td>
<td>Therapeutic use (TU)</td>
</tr>
<tr>
<td>Immunosuppressive Agents (D007166)</td>
<td>Therapeutic use (TU)</td>
</tr>
<tr>
<td>MAP Kinase Signaling System (D020935)</td>
<td>Drug effect (DE)</td>
</tr>
<tr>
<td>Plasmapheresis (D010956)</td>
<td></td>
</tr>
<tr>
<td>Interleukin-13 (D018793)</td>
<td>Antagonists &amp; inhibitors (AI)</td>
</tr>
</tbody>
</table>

Table 2. List of UMLS SN relationships between the semantic types Disease/Syndrome, Organism Function and Pharmacologic Substance.

<table>
<thead>
<tr>
<th>Subject/Object</th>
<th>Disease or Syndrome</th>
<th>Organism Function</th>
<th>Pharmacologic Substance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease or Syndrome</td>
<td>Associated_with, Co-occurs_with, Result_of, Degree_of, Process_of, Manifestation_of, Precedes, Affects, Occurs_in, Complicates</td>
<td>Result_of, Process_of, Manifestation_of, Affects</td>
<td></td>
</tr>
<tr>
<td>Organism Function</td>
<td>Process_of, Results_of, Affects</td>
<td>Co-occurs_with, Result_of, Degree_of, Process_of, Precedes, Affects</td>
<td></td>
</tr>
<tr>
<td>Pharmacologic Substance</td>
<td>Diagnoses, Treats, Complicates, Affects, Prevents, Causes</td>
<td>Complicates, Disrupts, Affects</td>
<td>Interacts_with</td>
</tr>
</tbody>
</table>

Table 3. Simplified sample record from the UMLS MRCOC file, containing the following fields: PubMed Unique Identifier (PMID); the dates related to the publication of the paper and its related MeSH indexing year; whether both MeSH terms are the main topics in the publication (ZY) or not (ZN); a description of the first MeSH term that consists of the unique identifier for MeSH heading term (MeSH DUI), the corresponding UMLS concept unique identifier (UMLS CUI) and the list of comma-separated MeSH subheadings that qualify the MeSH term; and an analogous description of the second MeSH term.

<table>
<thead>
<tr>
<th>PMID</th>
<th>Earliest year, pub date, article date, date completed, indexing year</th>
<th>Major Topics</th>
<th>MeSH Descriptor 1</th>
<th>MeSH Descriptor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>20278133</td>
<td>19461001</td>
<td>19461001</td>
<td>20100318</td>
<td>2010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This will be addressed by analyzing, in parallel, (i) patterns of MeSH term / subheading co-occurrence as extracted from MEDLINE metadata, and (ii) the text of the paper abstracts, using SemRep. In both cases, the target representations are triplet-based predications, using UMLS SN relations. Based on the frequency of natural language predicates in the abstracts, we attempt to infer relations that interpret the statistical associations regarding the distribution of 83 subheading types in MeSH annotations.
Rather than formal-ontological relations, we expect, primarily, to extract associative relations between the subject and the object concept, which express what is typical or probabilistic. We are not interested in statements like \(<\text{Nephrotic Syndrome} \mid \text{Is}_a \mid \text{Kidney Disease}>\) which are already extensively covered by ontologies like SNOMED CT or OBO (“all nephrotic syndromes are kidney diseases”). Instead, we focus on triples like \(<\text{Corticosteroids} \mid \text{Treats} \mid \text{Nephrotic Syndrome}>\), which represent contingent or probabilistic (“non-ontological”) knowledge, which cannot be translated into first order logics\(^{17}\) and is not expected to be found in domain ontologies. However, this is exactly the kind of content we consider more “interesting” from a biomedical knowledge representation point of view\(^{18}\).

We limited the scope of our experiment dataset to the MEDLINE records published in the last 5 years (2009-2013). Besides, we focused on the most relevant MeSH terms that (i) co-occur in the MEDLINE records and (ii) are linked to the UMLS SN types Disease/Syndrome and Pharmacologic Substance. Despite of such limitations, the amount of data involved is still huge and therefore requires efficient and scalable methods typical for big data analytics\(^{19}\), including a distributed framework that supports multithreaded, massively parallelized computing tasks.

**Material and Methods**

**Resources:** The main resource is the UMLS MRCOC file, which provides information about pairs of MeSH terms that co-occur in each MEDLINE record related to a specific MeSH indexing year. We used the 2014AA release of the detailed MRCOC version with a size of 131 GB. Each line of this file (Table 3) contains, among other data, a pair of co-occurring MeSH terms, the subheadings assigned to them, and the identifier of the MEDLINE record where these MeSH terms co-occur. The number of subheadings is variable, depending on the content of the article. Each MeSH term points to one UMLS unique concept identifier (CUI), which is linked to one or more UMLS semantic types in the files MRSTY and SRSTRE\(^{1}\).

**Tools:** SemRep is a system that analyses text and extracts semantic predications as triples. Their subjects and objects are represented as UMLS CUIs and the predicates correspond to one of the 54 UMLS SN relations. The output of SemRep is classified into three categories: TEXT, ENTITY and RELATION. TEXT describes the textual section analyzed by the system. ENTITY is related to a particular UMLS concept identified in the text, and RELATION describes a semantic relation between two entities found in the text.

**Methodology:** The creation of predications from MEDLINE metadata consists of five main phases: (i) aggregation of the co-occurring concepts and their MeSH subheading information; (ii) filtering of the aggregated co-occurrences based on their log-likelihood rates\(^{20}\) (LLRs); (iii) clustering of the co-occurring concepts based on the accumulated MeSH subheadings; (iv) extraction of semantic predications from MEDLINE records associated with the co-occurrences of every resulting cluster using SemRep; and (v) identification of the statistically significant predicates of each cluster.

- **The first phase** is the aggregation of co-occurring concept pairs in order to select the most relevant ones. The input data is the list of co-occurrences from MRCOC described in Table 3, and the output will be the list of aggregated PubMed identifiers, the MeSH and UMLS IDs of the first term with a list of aggregated MeSH subheadings, the MeSH and UMLS IDs of the second term with a list of aggregated MeSH subheadings (Figure 1). Due to the large data volume in the detailed version of MRCOC, we have applied the MapReduce\(^{21}\) programming paradigm and the Amazon cloud services (Amazon EMR\(^{22}\) and Amazon S3\(^{23}\)) together with Apache Hadoop\(^{24}\) for generating the aggregated version of MRCOC. MapReduce provides two types of procedures: MAP and REDUCE. The functionality of MAP is to filter and sort elements, where REDUCE takes the output of MAP and performs certain operations on the processed values. The data is represented as key/values pairs, thus facilitating data access and distribution across several computers. Hadoop is a software framework that supports the distributed processing of large data sets across clusters of computers. Consequently, processing the MRCOC file using both, MapReduce and Hadoop, can be easily scaled up and parallelized.
The second phase of our methodology is the calculation of log likelihood ratios (LLRs) and the co-occurrence filtering based on such rates. LLR reflects the statistical relevance of a pair of concepts regarding their percentage in the dataset and the percentage of other concept pairs. The required parameters to calculate the LLR of concept C1 that co-occurs with concept C2 are:

1. Number of co-occurrences of C1 and C2 (#C1\(\cap\)C2);
2. Number of co-occurrences of C1 without C2 (#C1\(\cap\neg\)C2);
3. Number of occurrences C2 without C1 (\(\neg\)C1\(\cap\)C2);
4. Number of co-occurrences where neither C1 nor C2 occur (#\(\neg\)C1\(\cap\neg\)C2).

These parameters were obtained during the aggregation of list of MeSH subheadings using Amazon cloud services. The formula in Figure 2 explains how the log-likelihood ratio is computed. The function H represents the Shannon entropy. H(matrix) indicates the entropy of #C1\(\cap\)C2, #C1\(\cap\neg\)C2, #\(\neg\)C1\(\cap\)C2 and #\(\neg\)C1\(\cap\neg\)C2. H(mRows) is the sum of entropies of the pairs <#C1\(\cap\)C2, #\(\neg\)C1\(\cap\)C2> and <#C1\(\cap\neg\)C2, #\(\neg\)C1\(\cap\)C2>. Finally, H(mCols) is the sum of entropies of the pairs <#C1\(\cap\)C2, #C1\(\cap\neg\)C2> and <#\(\neg\)C1\(\cap\)C2, #\(\neg\)C1\(\cap\neg\)C2>. In order to filter the less significant co-occurrences we applied a threshold of 10.83, which corresponds to the chi-squared test with one degree of freedom and a p-value < 0.001.

\[
LLR = 2 \times (H(\text{matrix}) - H(\text{mRows}) - H(\text{mCols}))
\]

The third phase is the co-occurrence clustering. The resulting list of co-occurrences from the previous phase is firstly, filtered by the semantic types of both concepts and, then, clustering them by the list of aggregated MeSH subheadings. Thus, we obtain different sets of co-occurrences for each pair of semantic types, such as Disease or Syndrome with Organism Function, or Pharmacologic Substance with Pharmacologic Substance. We hypothesize that each pair of semantic types can be interpreted as one or more UMLS SN relations. Their number can be used as input parameter to define the number of clusters to obtain during the clustering process. We use WEKA, which provides a set of clustering and machine learning algorithms, which can be applied to data mining tasks, from which we chose the k-means clustering algorithm. The variables to be used to cluster the co-occurrences are the list of aggregated MeSH subheadings. The number of occurrences of each subheading in the list is normalized by the total number of co-occurrences found for its related concept pair, and, therefore, the value for each subheading will be in the range [0, 1]. As there are 83 types of MeSH subheadings and because each concept in the co-occurrence has a different list of subheadings, the total number of features is 166.

In our experiments, we have focused on the combination of the semantic types Disease or Syndrome with Pharmacologic Substance, for which UMLS SN defines six predicates: Diagnoses, Affects, Treats,
Complicates, Prevents and Causes (Table 2). The number of suggested predicates provided by UMLS SN can be used as the input parameter of the number of clusters to be produced by the k-means algorithm. Moreover, we can also investigate the use of clustering algorithms that can estimate the best number of clusters.

- In the **fourth phase** of the methodology, predictions are extracted from MEDLINE abstracts using SemRep. However, the accuracy of such predications is limited by the ambiguity and complexity of natural language. Nevertheless, we can use the outcome of this analysis as indicators for the most relevant predicates for each co-occurring concept pair. Subject to analysis are the abstracts of those MEDLINE records that are related to the co-occurrences of the resulting clusters from the third phase. This phase yielded a list of predicates directly related to each pair of co-occurring concepts.

- The goal of the final **fifth phase** is the association of the most relevant predicate to each resulting cluster. To this end, we analyze the predications generated by SemRep. The analysis is focused on the study of the frequencies of the predicates in each cluster. According to our hypothesis, the concept co-occurrences grouped together in a cluster are related with the same type of predicate. Therefore, the predicate frequencies found by SemRep for those co-occurrences should be also statistically significant.

We compared the results between clustering into five, six, and seven different clusters with the k-means clustering algorithm. The assignation of the relevant predicates to each resulting cluster must be consistent with the predominant types of MeSH subheadings. Besides, the inference of unrelated predicates for the same co-occurrences will be analyzed, i.e. if in the extracted predications two concepts are associated with both Treats and Cause relations.

**Results**

The total number of aggregated co-occurrences from MRCOC is around 99,000. However, the input dataset for clustering only includes the resulting list of co-occurrences with LLR > 10.83, which reduces the number to 15,886 co-occurrences. Examples of co-occurrences with higher LLRs in our dataset are < HIV infection, AIDS Drugs >, < Hypertension, Antihypertensive Agents >, or < Grippe, Influenza vaccines >. Such co-occurrences with a high LLR are very frequent in MEDLINE records for the analyzed period.

The k-means algorithm produced centroids for each cluster and, hence, we can classify the co-occurrences into the cluster with the smallest Euclidean distance to its centroid. Nevertheless, co-occurrences can be ranked depending on their distance to the centroid of each cluster, as a consequence, we could compare the ranks of co-occurrences of each cluster and discover which ones are representative of more than one cluster.

The parameter of the number of clusters for the k-means algorithm was, firstly, obtained from the six relations suggested by UMLS SN between our selected semantic types; and, secondly, we used the expectation maximization (EM) algorithm\(^\text{27}\), which gave us an estimated number of five or seven clusters depending on the provided minimum standard deviation. Thus, we compute the clustering using five, six, and seven clusters. The clustering is based on the list of aggregated subheadings of each co-occurrence. Consequently, some subheadings are more predominant than others are. In Table 4, Table 5, and Table 6, we show the list of MeSH subheadings that have higher frequency in the co-occurrences for each cluster.

Because of the clustering, we obtained the lists of co-occurrences that belong to each generated cluster. The identifiers of the MEDLINE records, which are related to each co-occurrence, are also included in the MRCOC file. During the aggregation of the list of MeSH subheadings, the identifiers of such records were also collected and, therefore, we can use them to gather abstracts that are going to be processed with SemRep.

The collected corpus contains around 1,500 abstracts related the top co-occurrences of each cluster where, at most, 50 abstracts of the same co-occurring concept pair were collected. Each abstract corpus was processed by SemRep. The results were analyzed to extract the percentage of each predicate from the resulting triples where the subject and object are identical with co-occurring concepts of the cluster. However, this exact match occurred only in roughly 10% of the abstracts. Nevertheless, we assumed this subset reasonably representative for the whole. Thus, we normalize the frequency of each predicate in each cluster by dividing the resulting frequency by the total number of predicates that match the co-occurrences in a cluster. In particular, SemRep could identify through the different corpora the following predicates: Treats, Prevents, Affects, Causes, Associated with, Predisposes, Augments, and Disrupts. The predicates Predisposes, Augments, and Disrupts do not belong to UMLS SN, but they were proposed by SemRep. Figures 3 - 5 visualize the percentage of each predicate per cluster.
Table 4. List of the most relevant subheadings for each co-occurrence concept and for each generated cluster. The resulting clusters were produced by k-means algorithm with the input parameter of five clusters.

<table>
<thead>
<tr>
<th>Cluster 0</th>
<th>Subheadings 1st concept</th>
<th>Subheadings 2nd concept</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drug Therapy (DT)</td>
<td>Therapeutic Use (TU), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>Prevention and Control (PC), Immunology (IM)</td>
<td>Immunology (IM), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Drug Therapy (DT), Prevention and Control (PC)</td>
<td>Therapeutic Use (TU), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Metabolism (ME), Blood (BL)</td>
<td>Metabolism (ME), Blood (BL)</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Chemically Induced (CI)</td>
<td>Adverse Effects (AE), Therapeutic Use (TU), Administration and Dosage (AD)</td>
</tr>
</tbody>
</table>

Table 5. List of the most relevant subheadings for each co-occurrence concept and for each generated cluster. The resulting clusters were produced by k-means algorithm with the input parameter of six clusters.

<table>
<thead>
<tr>
<th>Cluster 0</th>
<th>Subheadings 1st concept</th>
<th>Subheadings 2nd concept</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drug Therapy (DT)</td>
<td>Therapeutic Use (TU), Adverse Effects (AE), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>Immunology (IM)</td>
<td>Immunology (IM), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Drug Therapy (DT), Complications (CO), Prevention and Control (PC)</td>
<td>Therapeutic Use (TU), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Metabolism (ME), Blood (BL)</td>
<td>Metabolism (ME), Blood (BL)</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Chemically Induced (CI)</td>
<td>Adverse Effects (AE), Therapeutic Use (TU), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>Drug Therapy (DT), Metabolism (ME), Pathology (PA)</td>
<td>Pharmacology (PD), Therapeutic Use (TU)</td>
</tr>
</tbody>
</table>

Table 6. List of the most relevant subheadings for each co-occurrence concept and for each generated cluster. The resulting clusters were produced by k-means algorithm with the input parameter of seven clusters.

<table>
<thead>
<tr>
<th>Cluster 0</th>
<th>Subheadings 1st concept</th>
<th>Subheadings 2nd concept</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drug Therapy (DT)</td>
<td>Therapeutic Use (TU)</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>Prevention and Control (PC), Immunology (IM)</td>
<td>Immunology (IM), Therapeutic Use (TU), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Blood (BL), Diagnosis (DI)</td>
<td>Blood (BL)</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Metabolism (ME)</td>
<td>Metabolism (ME)</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Chemically Induced (CI)</td>
<td>Adverse Effects (AE), Therapeutic Use (TU), Administration and Dosage (AD)</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>Drug Therapy (DT), Metabolism (ME), Pathology (PA)</td>
<td>Pharmacology (PD), Therapeutic Use (TU)</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>Drug Therapy (DT)</td>
<td>Therapeutic Use (TU), Adverse Effects (AE), Administration and Dosage (AD)</td>
</tr>
</tbody>
</table>
Figure 3. Percentage of relations that were extracted by SemRep within the five clusters generated using k-means.

Figure 4. Percentage of relations that were extracted by SemRep within the six clusters generated using k-means.

Figure 5. Percentage of relations that were extracted by SemRep within the seven clusters generated using k-means.
Table 7. List of the most relevant predicates for each cluster generated by 5 k-means, 6 k-means and 7 k-means.

<table>
<thead>
<tr>
<th>Cluster 0</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
<th>Cluster 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 k-means</td>
<td>Treats</td>
<td>Prevents</td>
<td>Treats</td>
<td>Treats</td>
<td>Causes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Associated with</td>
<td>Prevents</td>
<td>Associated with</td>
<td>Causes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 k-means</td>
<td>Treats</td>
<td>Treats</td>
<td>Prevents</td>
<td>Associated with</td>
<td>Causes</td>
<td>Affects</td>
</tr>
<tr>
<td></td>
<td>Prevents</td>
<td>Affects</td>
<td>Prevents</td>
<td>Associated with</td>
<td></td>
<td>Associated with</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Causes</td>
<td></td>
<td>Predisposes</td>
</tr>
<tr>
<td>7 k-means</td>
<td>Treats</td>
<td>Prevents</td>
<td>Treats</td>
<td>Associated with</td>
<td>Causes</td>
<td>Treats</td>
</tr>
<tr>
<td></td>
<td>Associated with</td>
<td></td>
<td>Associated with</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The analysis of the results of SemRep produces the association of the predicates with the corresponding clusters that are indicated in Table 7. These associations were obtained by selecting those predicates that, at least, were present in 25% of the total predications extracted by SemRep. Using Table 7 we could generate predications such as: <Nesiritide — Treats — Heart failure>, <BCG vaccine — Prevents — Tuberculosis>, <Estrogen — Associated with — Endometrioses>, <Erythropoietin — Affects — Ischemia> and <Nevirapine — Causes — Stevens-Johnson Syndrome>. Moreover, when two or more predicates were assigned to a cluster we could generate predications such as <Anticoagulants — Treats — Arterial Obstructive Disease> and <Anticoagulants — Prevents — Arterial Obstructive Disease>. A first inspection of 50 predications by an expert did not spot any clearly wrong predication. However, due to the high threshold used (p-value < 0.001), we assume that the recall is rather low. A systematic assessment is still outstanding and will be done in the next round after the inclusion of more semantic types and improvements of the matching process.

Discussion and Conclusions

This study is mostly descriptive and refrains from a detailed quantitative analysis, as it constitutes the first important milestone of a larger research project. The inspection of Fig. 3 - 5 seems to be only partly discriminative, but exhibit some interesting aspects that will guide our further work:

- Only a few relations exhibit a clear profile, viz. Treats, Prevents, and Causes. However, there is a cluster that includes in similar proportion Treats and Prevents, which is easily explained by the fact that there are substances used for both treatment and prevention.
- There is a clearly different profile of the relation Causes, because what causes a disease, is rather unlikely to be used in its prevention or treatment.
- Augments and Disrupts are too infrequent to allow any statement. Associated with, Predisposes, and Affects are very little discriminative. This is not surprising, because these relations are very vague. Affects may include causation, prevention, and treatment, Predisposes is difficult to interpret for disorder – substance associations, and Associated with could be as the most generic predicate which subsumes all of the other ones. This shows that the predicates generated by SemRep, but which are not present in the UMLS SN are not helpful for a clear semantic interpretation of the clusters.
- We hypothesize that other relations are more important to describe disorder – substance associations, especially the relation Diagnoses, which describes a substance that can be used to diagnose a disorder.
- The lack of discrimination may also be due to the genericity of many MeSH concepts such as Dermatologic Agent or Vaccines, for which there might be specializations for all predicates under scrutiny. This may lead to the decision to ignore too general concepts in the future refinement of the method.

The main advantage of selecting the most relevant predicates for each cluster is that a low p-value of the predicate could reduce the errors in the resulting set of predications obtained with NLP. For example, from the text: “Even conventional immunosuppressive agents, such as glucocorticoids and cyclosporine, directly affect podocyte structure and function, challenging the immune theory; of the pathogenesis of childhood nephrotic syndrome in which disease is caused by T cells.” SemRep extracts the predication <Immunosuppressive Agent — Causes —
Nephrotic Syndrome. It is not surprising that a relation extractor fails with syntactically complex sentences like this one. This explains why a certain background noise is unavoidable when relying on NLP tools.

Another limitation is that the scope of SemRep is the whole UMLS Metathesaurus and not only its MeSH subset. This explains the low rate of matching between the concept pairs. This could be improved by inferring predications between MeSH concepts from predications between UMLS Metathesaurus concepts in general by traversal of hierarchical links. This approach can increase the number of matching predications and, thus, provide a bigger dataset.

The comparison between the inferred predicates (Table 7) and the resulting clusters from the three clustering experiments (Table 4, Table 5, and Table 6) allows us to identify the combination of MeSH subheadings that are closely related to each particular relation. From our experiments we obtained: (1) Treats is related to co-occurrences that the disease term is annotated with Drug Therapy and the substance with Therapeutic Use; (2) Prevents is associated with co-occurrences that the disease term is annotated with Prevention and Control, and Immunology and the substance with Immunology, and Administration and Dosage; (3) Causes is related to the co-occurrences that the disease term has Chemically Induced subheading and the substance has Adverse Effects, Therapeutic Use, and Administrative and Dosage; and (4) Associated with and Treats are related to the co-occurrences that both the disease and substance have the MeSH subheadings Metabolism and Blood.

Finally, an evaluation of the generated predications would be necessary to rate their plausibility. It is obvious that predicates that represent more than 80% of the extracted predications in a cluster by SemRep are more plausible that those which are closer to 25%. Besides, the evaluation of the distance of co-occurrences to the cluster centroid might detect weaknesses that could guide the generation of predicates.

Acknowledgements

This paper was performed as a part of the BMFacts project (BMFacts: Knowledge acquisition for a biomedical fact repository), funded by the Austrian Science Fund (FWF): [M 1729-N15].

References

22. Amazon Web Services, Inc. Amazon Elastic MapReduce (Amazon EMR); 2015. Available from: http://aws.amazon.com/elasticmapreduce/?nc2=h_ls
23. Amazon Web Services, Inc. Amazon Simple Storage Service (Amazon S3); 2015. Available from: http://aws.amazon.com/s3/?nc2=h_ls
Intelligent Simulation Model To Facilitate EHR Training

Vishnu Mohan, MD, MBI1, Gretchen Scholl, BS1, Jeffrey A. Gold, MD1
1Oregon Health & Science University, Portland, OR

Abstract

Despite the rapid growth of EHR use, there are currently no standardized protocols for EHR training. A simulation EHR environment may offer significant advantages with respect to EHR training, but optimizing the training paradigm requires careful consideration of the simulation model itself, and how it is to be deployed during training. In this paper, we propose Six Principles that are EHR-agnostic and provide the framework for the development of an intelligent simulation model that can optimize EHR training by replicating real-world clinical conditions and appropriate cognitive loads.

Introduction

The growth of electronic health records (EHRs) in the United States in the last decade has been meteoric, primarily as a consequence of the Health Information Technology for Economic and Clinical Health Act (the HITECH Act), enacted under Title XIII of the American Recovery and Reinvestment Act of 2009. The reimbursements afforded to institutions and individuals as a consequence of “meaningful use” regulations have also added substantially to the impetus of EHR adoption1.

As EHR use has grown, the potential for improvement of healthcare delivery quality has also increased2, but so has the potential for causing patient harm associated with EHR use3,4. This phenomenon can be attributed to a number of reasons, primarily because the addition of technological innovations into complex healthcare systems appear to be associated with a new breed of errors, a process which has been termed “technological iatrogenesis”5.

EHR training plays a critical role not only in fostering end-user EHR adoption, but also in their ability to efficiently use the EHR while delivering clinical care6. This realization led the Office of the National Coordinator for Health Information technology (ONC) to endorse the creation of curricular materials for entry-level HIT professionals that included hands-on learning with an EHR7. However, a standardized learning plan is not available to healthcare organizations that are implementing EHRs. Further, different EHRs, and even different implementations of the same EHR, tend to have variations in their interface, which makes the process of finding and reviewing information more difficult for end-users.

Typically, clinicians using the EHR need to find relevant information in a rapid fashion, often while they are multitasking, and operating in an environment that can be distracting8. Training clinicians to operate efficiently in such an environment while using the EHR in a manner that optimally facilitates medical decision making and patient care requires the development of EHR training protocols that best aid learning. Further, since the time clinicians spend in EHR training is in essence time they cannot devote to direct patient care, optimum EHR use training needs not only to be comprehensive but also efficient with respect to the use of clinicians’ time. Training must take into account the needs of all of the varying professional groups which interact with the patient. Finally, EHR training programs also need to consider the unique characteristics and assumptions of the community of practice they serve9, the degree of comfort that each individual clinician has with the EHR, and their baseline proficiency.

Many models of EHR proficiency training have been proposed, but the most commonly used formats are formal classroom/lab, and one-on-one/at-the-elbow training10. Most use a training or simulated EHR environment, since the use of real-life situations to impart EHR training has critical drawbacks, including the fact that imparting training in...
a production environment can introduce errors into a real patient’s chart, that no two real life cases are identical (i.e. the lack of replicability), and that training using actual patient records may violate Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rules. An increasing body of literature describes the potential of simulation training to improve acquisition of clinical skills, training of health care teams to effectively navigate complex systems and improve teamwork, communication and shared decision making. While the technology to improve the fidelity (realism) of simulation activities has improved dramatically, there are still significant barriers and limitations to incorporating the EHR into clinician training regimens. Firstly simulation environments are often very different than their production counterparts, marginalizing their utility when provider knowledge and use of the actual operational clinical information system need to be emphasized. Second, EHR training environments are not populated with actual patient information (in deference to HIPAA privacy rules) and typically contain just a few cases populated with sparse data. Typical cases within EHR training environments rarely recapitulate the complexity and specific clinical information encountered by end-users providing clinical care within the scope of their training, and rarely replicate the level of data density encountered in real-world utilization. Third, EHR simulations are typically not associated with the “cause-and-effect” real-time cascade of events that is seen in real life clinical settings – for example ordering an antibiotic for a febrile patient with an infection results in the normalization of their temperature and a change in their clinical status which results in the discontinuation of the antibiotic after an appropriate period of time.

In this paper, we describe a model for developing and building realistic, clinically complex simulations that can be used for optimizing end-user EHR training, in a manner that facilitates the simulation of realistic clinical cognitive loads as end-users participate in simulated activities.

The case for simulation

The primary advantage of using simulations for training is that they allow learning activities to be absolved of any risk of harming real patients. Simulation tools and technologies have advanced dramatically over years, and now allow both the training of classical procedural skills such as airway management of trauma patients in the prehospital setting\textsuperscript{11} or advanced cardiac life support (ACLS)\textsuperscript{12}, as well as more innovative learning, such as the use of simulation in teaching clinical judgment skills\textsuperscript{13} or in disease management and prevention\textsuperscript{14}.

By replicating complex real-life situations\textsuperscript{15}, simulations allow the evaluation of clinical skills and competence in a comprehensive manner\textsuperscript{16}. Further, we have used simulations to assess EHR safety\textsuperscript{17} and demonstrated that participation in EHR simulation helps clinicians better recognize patient safety issues\textsuperscript{18}.

Characteristics of an optimal simulation-based EHR training protocol

However there is still a need to develop an optimal EHR training protocol that utilizes simulation effectively. There are three primary barriers to optimal simulation use at present which will need to be overcome. First, default “training” cases included with most EHR distributions are overly simplistic, and contain sparse data. This is very different than real life patient records, which are often complex and contain a plethora of data. For example, a patient admitted to a typical ICU may generate more than 1,500 data points each day they are in critical care\textsuperscript{19}. Optimally training an ICU physician would require the use of simulation cases that replicate at least a significant proportion of the data complexity that they would be expected to encounter when they are providing direct patient care.

Second, EHR training should also be able to replicate the cognitive load that clinicians are subjected to when they peruse the patient chart. These will allow the clinician to understand and mitigate the errors of cognitive that are closely associated with their task load and EHR use patterns\textsuperscript{20}. For example, while EHRs can reduce errors associated with physician order entry, they may also paradoxically increase them\textsuperscript{21, 22}. Typical EHR training cases focus on entering one, or at most a few orders into the system in a setting where cognitive load is minimized, thus
minimizing the likelihood of any errors occurring during the training process, and missing a rich opportunity to emphasize a specific training objective that would allow end-users to identify, correct and mitigate a common potential cause for patient harm.

Third, the current EHR training paradigm continues to focus on the structure of the EHR as the prime anchor for training, rather than the function of the EHR within the context of the cognitive tasks the clinician needs to perform using the EHR. A clinical simulation scenario is better suited to contextualize the use of the EHR in clinical settings than the traditional classroom approach where a specific topic of EHR functionality is first identified, followed by a demonstration of the task(s) required to achieve the desired functionality. By migrating away from a “learn-by-rote” model of EHR training towards “case-based” model of EHR training, learning objectives are better reinforced and clinicians are able to translate the concepts they have learned in their EHR training session much more effectively to their clinical practice.

**Building the infrastructure**

In order to achieve an effective training paradigm, the EHR simulation environment itself should be capable of achieving the necessary degree of fidelity that replicates the system in the clinical setting. To accomplish this in an optimal fashion, one would employ a simulation version of the clinical EHR that would allow for full utilization of system functionality without risking either breach of patients' protected health information or potentially impacting other connected systems (e.g. pharmacy or billing clinical information systems).

In our experience, the simulated instance of the EHR should contain the following characteristics: first, it should be routinely “cloned” from the existing production environment (with all patient identifiable information deleted) thus allowing users to have the most up to date version of the EHR, as well replicate the individual provider's user-specific customizations, macros and order sets. Second, the system must be populated with simulated medical records that have been designed specifically for the workflow being assessed, with cases that reproduce the clinical complexity and scenarios encountered in actual practice. Third, cases must be temporally aligned with the proposed exercise. Specifically, complex cases need to be carefully designed and built into the system in advance of any training exercise. Rather than populate the simulation environment with patient data that exists in the distant past (as perceived by the end-user), it is essential that patient data, if appropriate, be carried forward to the day of training. Finally, real time report results need to be generated directly into the EHR rather than be verbally communicated to the end-user during the simulation activity. Ensuring this feature is often the most difficult; depending on the exercise, results may need to be customized “in the moment” and informed by the performance of trainees. However, this functionality is essential for replicating realistic EHR use patterns in high-stress situations, especially when considering end-user information processing, as well as subsequent diagnostic reasoning and shared decision making.

Once the simulation environment of the EHR is created, it must be deployed in a manner similar to its clinical use. When incorporating the EHR into other simulation activities, it is important to ensure that the simulation workstations resemble, at least in terms of form and function, the actual workstations used in the clinical environment. This includes replicating monitor size, the use of mobile workstations and ergonomics similar to the clinical care environment when applicable. This is important to not only eliminate potential confounders introduced by unfamiliar hardware, but also fosters cognitive thinking in a fashion that mimics real life conditions. Ideally, once established the actual simulation center itself could be EHR agnostic, allowing multiple EHRs to be utilized in simulations depending upon the needs of the organization, the tasks at hand and the learners themselves.

**Conceptualizing an Intelligent Simulation Model**

Engaging in simulation activities allows learners to move from the knowledge or comprehension levels of Blooms taxonomy to application and analysis. The provision of high fidelity simulation that replicates a level of cognitive
load comparable to real life clinical situations may also allow learners to synthesize knowledge for subsequent application in real life patient care.

Simulation activities promote learning without endangering patient safety, utilizing standardized replicable scenarios. High fidelity simulations allow learners to engage in scenarios that are closely reminiscent of real life situations, and thus engage in analogical reasoning that integrates new information learned during the simulation to prior accrued experiential clinical knowledge, thus improving competency.

We have previously described a model for intelligent case design that emphasized a patient-centric record and utilized an interprofessional team-based approach to designing complex cases. Additionally we have also described a framework that allowed us to test how high fidelity simulations can be utilized to test new EHR training models or interface designs in a manner that facilitates error recognition and improves patient safety. We built upon this framework and utilized intelligent cases to facilitate a model that allowed the generation of realistic cognitive loads as learners participated in the simulation activity. The intelligent cases designed by us specifically allowed end-users to identify both normal as well as abnormal patient data by recognizing low and high-end cutoffs in the context of the patient’s clinical presentation rather than depending on the traditional defined range. Our cases included both common as well as rare data, at a level of density that reflected real-life clinical situations. The data also allowed end-users to not only find relevant data, but also to recognize when relevant data was missing from the user interface screens they were expected to use. And finally, our cases incorporated best practice and meaningful use criteria to emphasize the need to incorporate standards of care into efficient EHR use.

Deploying cases designed using the principles delineated above allows end-users to be trained using a model that closely replicates the cognitive load they encounter while taking care of patients, and allows them to anchor simulation-based EHR training in the context of providing clinical care. However these cases are best deployed using a simulation paradigm for EHR training that is also based on intelligent design principles.

**Building intelligent simulations**

We propose Six Principles for effective Intelligent Simulations that foster effective EHR training. These Principles emphasize the need to provide a training experience that closely mimics real-life cognitive and decision-making paradigms, and allow EHR training to occur in the context of the clinical care that the end-user is trained to provide. They are independent of the EHR being used to conduct the simulation.

1. **EHR training should be conducted in the form of a simulation exercise that describes a clinical scenario familiar to the learner.** The training session should be anchored utilizing the clinical context of the simulation, rather than on specific elements of EHR functionality. The simulation exercise should focus on pre-determined optimal clinical outcomes that define efficient EHR use, rather than the EHR-specific tasks themselves. For example, if the simulation trains the end-user to learn order entry functionality for prescribing medications, the clinical context (for example a patient develops an abnormal heart rhythm necessitating the prescription of a specific medication to counter it) should drive the training session, and success should be aligned to the end-user's ability to visualize and examine specific information within the EHR, rather than focus only on the EHR itself -- in the case of the example noted above, one success measure could be the ability of the end-user to visualize and recognize the patient's changes in clinical and vital signs after prescribing the appropriate medication.

2. **Simulations should replicate real life processes.** This implies that cases should not be too simplistic (as is the case with most current training cases) or divorced from reality. Clinicians should be immersed in the environment as much as the simulation infrastructure allows, and cognitive skills utilized in real-life settings should be brought to bear as the simulation training progresses.

3. **The simulation should use a standardized case structure and format.** Most importantly, these cases should replicate both the clinical complexity as well as the workflows of the clinician being trained, with adequate data...
density and complexity. In addition, two vital characteristics of real-life cases need to be included within the intelligent simulation. First, the structure of simulation cases should replicate real-life conditions. Clinical scenarios should be plausible, and documentation within the patients chart (such as clinical notes, X-rays or laboratory tests) should reflect the clinical condition in a manner that replicates real-life considerations. Second, the diagnostic reasoning processes that the trainee brings to bear should mimic real-world cases, thus attempting to maintain intellectual fidelity in addition to data fidelity.

[4] Each simulation needs to address the routine “cause-and-effect” phenomenon that characterizes EHR use in the clinical setting, specifically with respect to the temporal sequencing of data. Typically, training EHR cases showcase events that have occurred in the past (typically on the day the cases were incorporated into the instance). This creates an artificial temporal sequence that end-users usually find unrealistic. By utilizing “rolling” cases that allow patient data to be accessed as they would in real life (i.e. yesterday’s lab values can be accessed by the end-user scrolling back by one day, and last week’s vitals by going back a week in time) the cognitive load placed on the learner is realistic and sustained. Complex cases will need careful design to reflect this, and patient data will need to be routinely carried forward up to the day of testing. End-users will also need to access real-time reports as they engage in training. This allows appropriate information processing and use of the EHR in a manner that fosters diagnostic reasoning.

[5] Simulations should be conducted using an instance of the EHR that has been replicated from the production environment. The training instance should incorporate all the individual customizations applicable to the end-user receiving training, and also allow the use of protocols (for example logins, macros) that are identical to those utilized in the production environment. This allows end-users to learn using the same screens and customized features that they are familiar with when they take care of patients, thus optimizing their level of learning with the EHR.

[6] Simulation training protocols should be developed by an interprofessional team. The practice of medicine today is clearly a team activity. We believe that the perspectives of members of an interprofessional team are invaluable in developing simulation cases and protocols for specific disciplines of clinicians. For example, a nurse may have insight that allows a more sophisticated EHR training protocol to be defined for a group of physicians, or a pharmacist may contribute significantly to the EHR training of nurses. Involving multiple disciplines in the development of EHR training protocols also fosters collaboration, and strengthens patient safety.

A summary of the six design principles delineating the focus and evaluation paradigm associated with each individual principle is presented in Table 1.

**Tailoring simulations**

Intelligent simulation designs incorporate clinical scenarios that replicate the cognitive load that clinicians experience when they are actively utilizing the EHR to provide patient care. This requires not only calibrating the amount and complexity of data presented during the simulation, but also customizing the simulation to the specific role of the clinician. For example, a case targeted to physicians may need additional customization to ensure that it retains adequate high-fidelity and cognitive complexity for nurses, pharmacists on other non-physician clinicians who may participate in simulation activities. Cases will need to be tested to ensure that pertinent and appropriate data elements are displayed not only on physician EHR screens, but also on role-specific EHR screens for any non-physician clinicians who may participate in the activity.

Further, the level of cognitive load that can be assumed by individual clinicians is also dependent on their clinical competency, and experiential factors such as their level of training. A medical student participating in a high fidelity EHR simulation assimilates and processes clinical data differently than an attending physician. Developing
intelligent simulation requires designers to acknowledge the importance of calibrating both the complexity and the cognitive load of the simulation to match the level of competency and experiential learning of individual subjects.

<table>
<thead>
<tr>
<th>Design Principle</th>
<th>Focus Of Simulation Activity</th>
<th>Simulation Evaluation Paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorporate EHR training elements within the contextual framework of the clinical simulation activity</td>
<td>Simulation activity should be anchored using clinical aspects of the scenario, as opposed to EHR functionality</td>
<td>The clinical context should inform EHR use and data retrieval by the subject, with evaluation measures driven by the clinical context as opposed to EHR design-related elements</td>
</tr>
<tr>
<td>Replicate real life processes with adequate levels of calibration</td>
<td>Simulation activity should reflect a high level of realism and be framed in context that is appropriate for acuity, level of care, and clinical cognitive load</td>
<td>High fidelity simulation mandates immersion of the clinician in the simulation, to the best extent possible within the constraints of the environment</td>
</tr>
<tr>
<td>Utilize a standardized case structure and format, and duplicate real-life clinical workflows</td>
<td>Simulation activity should reflect appropriate level of data complexity and density, feature plausible clinical scenarios, and replicate the cognitive load encountered by clinicians in real life clinical care settings</td>
<td>Data complexity and density, both with respect to individual data elements as well as trends, should reflect the level of complexity expected from the clinical scenario. Documentation and data can be reviewed by clinical subject matter experts to ensure that they maintain both data and intellectual fidelity</td>
</tr>
<tr>
<td>Allow optimum cognitive reasoning by facilitating the &quot;cause-and-effect&quot; phenomenon of EHR data retrieval</td>
<td>Simulation activity case design should focus on ensuring that data is appropriately sequenced temporally, to facilitate recognition of trends. Case designers should also incorporate real-time reports into the simulation activity to foster realistic diagnostic reasoning</td>
<td>Case testing prior to conducting simulation activity should confirm that patient data is available to subjects in a fashion that mimics real life, i.e. with appropriate adherence to temporal sequencing pertinent to the clinical scenario</td>
</tr>
<tr>
<td>Replicate simulation instance from production environment</td>
<td>Simulation environment should closely duplicate the production equivalent. Individual clinician customizations should be brought forward from production to simulation environments</td>
<td>The simulation environment should be audited to ensure individual end-user customizations (logins, documentation macros, order sets, institution or department-specific clinical decision support rules) etc.) have been replicated prior to initiation of simulation activities</td>
</tr>
<tr>
<td>Develop simulations utilizing an interprofessional team</td>
<td>Interprofessional clinical perspectives should be actively solicited while building cases for simulation activities</td>
<td>The team building intelligent simulation cases should comprise of members from different clinical disciplines that truly reflects the interprofessional constitution of the team delivering patient care in a similar clinical real-life context</td>
</tr>
</tbody>
</table>
Conclusion

Using simulation infrastructure deliver end-user EHR training can optimize learning, clinician efficiency and promote patient safety. Rather than continuing to utilize current training models, developing a strategy that incorporates intelligent simulation within clinician EHR training programs has the potential to improve how clinicians use the EHR, reduce patient harm, and promote superior clinical quality of care.

References

Developing the Pathologists’ Monthly Assignment Schedule: A Case Study at the Division of Anatomical Pathology of The Ottawa Hospital

Amine Montazeri, MSc, Jonathan Patrick, PhD, Wojtek Michalowski, PhD, Diponkar Banerjee, MBChb, FRCP, PhD

1 Telfer School of Management, University of Ottawa, Ottawa, ON; 2 Division of Anatomical Pathology, The Ottawa Hospital, Ottawa, ON

Abstract
In the Division of Anatomical Pathology of a teaching hospital at the beginning of each month, clinical managers assign expected daily pathology requests to the pathologists on duty. Since the number of these requests is usually large and a division employs a number of pathologists with different sub-specialties, the size of the problem is significant and finding a feasible assignment schedule manually is time-consuming. Moreover, every time there is a need to change, a new assignment schedule needs to be developed taking into account all the pre-defined constraints including pathologists’ availability, sub-specialty mix, teaching/research releases, etc. In this paper we describe an analytics optimization model embedded in a decision support tool that helps the clinical managers of the division determine the optimal monthly assignment schedule. The decision support tool has been validated using data from the Division of Anatomical Pathology at The Ottawa Hospital in Ottawa, Ontario, Canada.

Introduction and Problem Statement
Developing an assignment schedule for the pathology laboratory at a teaching hospital is a process that assigns the individual pathologists to pathology requests taking into account pathology sub-specialties, the volume of requests, and any teaching/research functions that impacts each pathologist’s availability. As the complexity of clinical pathology grows, the time spent on creating the pathologists’ assignment schedule increases creating a need for a computer-aided scheduling support tool. In this paper we describe a support tool that was developed in collaboration with the clinical managers and pathologists from the Division of Anatomical Pathology (DAP) at The Ottawa Hospital (TOH). TOH is a teaching hospital affiliated with the University of Ottawa that provides inpatient and outpatient services on three campuses located in different parts of the city. The DAP is located on one of these campuses (General Hospital campus) and serves the entire TOH. It also receives pathology requests from the community hospitals and from private pathology laboratories serving the City of Ottawa and surrounding areas. Each day about 200 pathology requests of differing complexity arrive from the operating rooms of TOH, the clinics affiliated with the hospital, the community hospitals, and from small laboratories. These requests are recoded as the “cases”, with each case involving the analysis of one or more specimens that are further subdivided into slides. In this paper we are concerned with the assignment schedule at the level of the specimen. We do not describe the process of creating the specimens from a tissue but want to stress that each specimen may include (depending on its type) anywhere from 3 to over 200 slides that need to be examined by a pathologist to whom a particular specimen is assigned. The 36 full time pathologists working at the DAP are salaried and their responsibilities, apart from clinical work include teaching and research that impact their availability in a given day, week, and month. These pathologists cover services spanning 26 sub-specialties (i.e. liver, breast, neurology, etc.).

In order to process all pathology requests, at the beginning of each month clinical managers develop the assignment schedule for the following month. This schedule assigns each type of specimen to a pathologist for each working day of the month. It must take into account the following requirements:

- **Pathologist’s sub-specialty**: Each pathologist can assess only those specimens that are within his/her expertise such as breast, head and neck, musculoskeletal, to name a few.
- **Pathologist’s availability**: Due to a number of clinical and non-clinical responsibilities, each pathologist’s availability needs to be assessed on daily and weekly basis. We measure availability in a given week using “full time equivalent (FTE)” fraction (i.e. for a pathologist working full time in a given week FTE = 1).
The service weight: The amount of time it will take to assess all slides of a given specimen type in a given day is clearly a stochastic variable. However, these times are estimated using pre-determined service weights that are different for each specimen (i.e., a service weight of 2.0 for neuropathology indicates that the expected workload on a given day for neuropathology should require the services of two pathologists). There is no agreement in the literature (and practice) as to how the service weights ought to be determined, but this discussion is beyond the scope of our research. Clinical managers at the DAP rely on weights that are based on the L4E indicator used in a number of North American laboratories to determine pathologists’ workload.

The above requirements define what we call “hard requirements” as they need to be satisfied by any assignment schedule. However, there is a number of “soft requirements” that should be considered when a reasonable assignment schedule is created and these requirements include:

- Consistent assignment in a week: It is desirable that within a given week a pathologist is assigned to work on specimens belonging to a single sub-specialty.
- Rotation of the specimens: In order to maintain the clinical skills required for the analysis of a given sub-specialty, each pathologist should regularly rotate through his/her areas of expertise.
- Prioritization of the specimens: Due to vacations and other absences, it is not always possible to cover every sub-specialty on every day. Faced with this dilemma, the DAP prioritizes between subspecialties in order to ensure that the crucial ones are covered. For instance, full coverage requires three pathologists to be assigned to a given sub-specialty. Faced with insufficient resources, the DAP will choose to reduce that coverage to two pathologists as a first step. This ranking of coverage priorities needs to be respected.

Currently the clinical managers at the DAP use a manual approach to scheduling that relies on data recorded in a spreadsheet. Since the size of the pathologists’ assignment schedule is large, finding a satisfactory assignment manually is time-consuming and can take a number of iterations over a number of days to complete. Moreover, every time there is a need to revise the schedule the process must start over again.

In this paper we describe a computer-based decision support tool for developing an assignment schedule. This tool has two main components – an analytics optimization model that is used to develop the optimal assignment schedule taking into account both the hard and soft requirements and a spreadsheet-driven interface that is similar to what is being used now in the division and allows the managers to manipulate and revise the assignment schedule in order to assess a number of scheduling scenarios or consider special cases.

The next section presents a brief review of the research on using the assignment problem in healthcare. This is followed by a description of the analytics model that we have developed. The decision support tool is described in the Implementation section. The comparison of the assignment schedules developed automatically with those used by the division is presented in the Results section. The paper ends with some concluding thoughts.

Related Work

The development of an assignment schedule is an important analytics problem that aims to find an optimal allocation of n available staff members to m positions or tasks in a system. During the last few decades, researchers have developed a number of assignment models for various settings such as industrial systems, educational institutions and healthcare organizations all designed to help managers allocate tasks to resources (including human resources and equipment). A number of assignment models, specific to healthcare, have been developed. The type of model developed depends on the type of organization and the specific characteristics of the problem to be addressed. Assigning nurses and physicians (mostly in the Emergency Department) to shifts so that staffing requirements are met represents one of the most common problems and the analytics models include stochastic programming, goal programming and genetic programming models. Home care staff assignments require the development of an assignment schedule connecting health providers with patients based on factors such as workload restrictions, provider qualifications and preferences, acuity levels of patients, an overtime penalty and staff satisfaction. The majority of the home care assignment problems use optimization models.

To the best of our knowledge, our work is the first attempt to apply analytics to develop the optimal assignment schedule in a pathology department. The proposed model is similar to those used elsewhere but has a number of
distinct features: it deals with a situation where each pathologist has multiple sub-specialties and explicitly considers
the need for pathologists to rotate through the types of specimens in order to maintain the required skills in a given
clinical sub-specialty. It is worth mentioning that there are commercial scheduling systems designed to develop an
assignment schedule for the pathologists. However, they are mostly concerned with implementing the existing
scheduling practice in terms of rules rather than developing the kind of optimal assignment schedule provided by the
system described in this paper. A good example of such assignment is Q-Genda’s pathologist scheduling software
(Q-Genda corporation) that allows the managers to develop different scheduling scenarios according to a set of
decision rules. However, because of the way that these scenarios are created, it is not possible to determine if
proposed assignment is truly the best one or if it just meets basic requirements.

Methods

The development of the pathologists’ assignment schedule is a decision-making problem that involves the analysis
and assessment of a number of alternative schedules each with different characteristics. This type of problem is
solved using analytics methods that allow the user not only to assess possible alternatives very quickly but also to
identify the optimal one based on the measurable criteria. In this paper we describe the use of an analytics approach
that involves the development of a mixed-integer optimization model that incorporates both hard and soft
requirements. This model has the following four components: decision variables, model parameters, objective
function, and constraints. Each component is described in greater detail below and the full model is provided in the
Appendix.

Decision variables: The decision variables represent the unknown (to be determined by the model) daily
assignments of the pathologists to specimens on each day. For the DAP, assuming that a monthly assignment
schedule needs to be developed for 20 business days on average (DAP does not work on the weekends and statutory
holidays), this implies that there are 18720 possible daily assignment combinations. However, the incorporation of
the soft constraints requires the creation of auxiliary decision variables bringing the total to 43649.

Model parameters: Every model requires known and measurable inputs that are called “parameters”. We use five
types of parameters:

- The FTE for each pathologist working full time has a value of 1 while for the part timer it can be any
  number between 0 and 1 (i.e. for a pathologist working two out of five days a week the value of his/her
  FTE is 0.4). Values for the FTE fraction parameters were obtained from the DAP.
- The service weights for each type of specimen were determined by the DAP clinical managers as explained
  earlier.
- The availability of a pathologist on a given day is captured by a binary parameter (value 1 if the
  pathologist is available and 0 if the pathologist is unavailable). These values change from month to month
  and reflect each pathologist’s workload taking into account teaching/research obligations, holidays, etc.
- Another binary parameter captures which specimens are within each pathologist’s area of expertise. (When
  a pathologist is able to diagnose a given specimen type then the value of this parameter is 1. It is 0
  otherwise). This information was provided by the DAP.

Objective function: The purpose of the objective function is to establish a performance measure for assignment
schedules in order to derive the optimal solution. In our model the objective function is created by amalgamating the
soft requirements outlined earlier. Thus, the objective function has three parts, each with a specific weight of
importance that can be assigned by a clinical manager. In the case study described later in the paper the values of
the weights were determined experimentally so the resulting assignment schedule best matches the expectations of
the clinical managers.

Constraints: The constraints describe in a formal way the constraining factors that limit the development of an
assignment schedule. Model constraints include hard requirements as well as a number of additional constraints
related to the need to produce feasible daily assignments over 20 days. The model we developed for the DAP has
52057 constraints excluding binary and non-negativity constraints imposed on the decision variables.

The model was solved using variant of Simplex algorithm implemented in IBM ILOG CPLEX Optimization Studio
running on a Dell T7600 desktop computer with Windows7. On average it took seven to ten minutes to obtain the
optimal assignment schedule for a given month.
Implementation

Responding to the clinical managers request that the user interface to the model should have a “spreadsheet – like” feel, we embedded the model within a customized Microsoft Excel platform that provides a user-friendly interface. The resulting scheduling decision support tool was named the Automatic Pathologists’ Scheduler (APS) system. It integrates the analytics model and the Microsoft Excel spreadsheet using a number of macros. The APS system has a hierarchical structure with a main menu (see Figure 1) that allows the clinical manager to access the different functions of the system.

In order to use the APS system, the clinical manager needs to enter the values of the parameters required by the analytics model. These parameters can be categorized into:

- **Dynamic parameters**: the values of these parameters have to be updated every month before solving the model. They include information about the holidays in a given month, pathologists’ availability and specific “hard-wired” assignments that task a given pathologist to a specific specimen type on a given day.
- **Static parameters**: the values of these parameters are updated infrequently (or do not need to be updated at all). They include the roster of pathologists working in a division, the list of sub-specialties, the FTE fractions and the service weights.

The APS system has the built-in functionality that allows for extended editing and control of the development of assignment schedules. This functionality includes:

- **Editing**: revising values of all parameters, including anticipated daily pathology requests. This function also allows for the manual development of a partial assignment schedule that is later used by the analytics model as a starting point, as needed.
- **Optimization control**: The user interface allows the manager to experiment with different values for the weights assigned to the components of the objective function. The purpose of this functionality is to provide the clinical manager with the ability to develop different assignment schedule scenarios depending on the changing importance of each of the soft requirements.
- **Schedule analysis**: A separate spreadsheet interface for analysing the schedule allows the clinical managers to evaluate and revise the schedule, if needed.

The output produced by the APS system is illustrated on Figure 2 (due to space limitations only a fragment of the spreadsheet is presented). The assignment schedule is organized by the days in a month (columns), types of the specimens (coded with the abbreviations used in the DAP and represented as the rows), and labels associated with individual pathologists (cell values). In the original schedule the cell values include pathologists’ initials, but for privacy reasons they are replaced here with P# labels. The dash (-) on the schedule indicates a non-business day. As an example pathologist P14 is assigned all cardiac specimens from April 7 to April 10. Starting on April 13 this type of specimen is re-assigned to pathologist P5 (until April 17, inclusive). In the schedule, two separate rows have been allocated to breast specimens (indicated by “Breast1” and “Breast2” labels) as historically two pathologists are required to analyze all daily breast specimen requests.
Results

The system’s performance has been validated by the clinical managers at the DAP and also by comparing manually developed assignment schedules with those created by the model. The results show that the APS system is effective in developing a schedule that meets most of the pre-defined scheduling criteria and as such represents a significant step forward in managing the pathologists’ scheduling problem. To illustrate, below we present and compare the assignment schedules generated by the APS system for the months of October and November 2014 against the schedules that were developed manually for those same months and implemented in the DAP. When conducting these comparisons, following the advice of the clinical managers, we looked at:

- Percentage of unassigned specimen types: in the schedule developed by the APS system, higher priority sub-specialties are assigned to the pathologists first. This may result in some lower priority sub-specialties being left unassigned due to an insufficient number of pathologists available. This measure is calculated as the percentage of all unassigned specimen types within a month in relation to all the assignments in this month.

- Percentage of inconsistent assignments: this measures the consistency of assignments within a given week. It is calculated as the percentage of all assignments that are not consistent within a week in relation to all the assignments in this week. An assignment schedule that has a lower value for this measure is preferred.

- Percentage of missed rotations: each pathologist ought to cycle through all sub-specialties within his/her area of expertise. This is accomplished by rotating what specimen type a pathologist is covering each week. Missed rotations (in %) are the number of sub-specialties within each pathologist’s area of expertise that is not covered in a given month divided by the total number of sub-specialties for each pathologist. This metric is aggregated over all pathologists to derive a single measure. It is important to note that it may not be possible to cover all sub-specialties in a given pathologist’s area of expertise while still maintaining consistency with the week as some pathologists have 5 or more sub-specialties. However, an assignment schedule that has less missed rotations is clearly better.

Looking at the results presented in Table 1, we note that the APS system consistently outperforms the manually developed assignment schedule when considering the percentage of inconsistent assignments. The differences between the APS and the manual schedule on the other measures are negligible indicating that the model quite accurately reflects current practice.

Table 1. Comparison of automatic and manual assignment schedules for October and November 2014.

<table>
<thead>
<tr>
<th>Period</th>
<th>Unassigned specimen types (%)</th>
<th>Inconsistent assignments (%)</th>
<th>Missed rotation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>APS system</td>
<td>Manual assignment schedule</td>
<td>APS system</td>
</tr>
<tr>
<td>October 2014</td>
<td>1.5%</td>
<td>0.8%</td>
<td>14%</td>
</tr>
<tr>
<td>November 2014</td>
<td>0.6%</td>
<td>0.1%</td>
<td>12%</td>
</tr>
</tbody>
</table>
Discussion

In this paper we describe the APS system that helps clinical managers of the DAP to develop a monthly assignment schedule. The system creates the schedule by solving an analytics model that takes into account all hard and soft scheduling requirements considered currently by the managers. Providing the clinical managers with such an easy to use support tool has value for a number of reasons. First, the APS system helps to create the optimal assignment schedule very quickly. Secondly, the proposed schedule can be used by the managers to either make additional adjustments taking into account intrinsic requirements or to create a number of scenarios considering different staffing possibilities, specimens’ volume levels, etc. Finally, the APS system gives flexibility in establishing and revising all model parameters to reflect for example a sudden change in the pathologists’ availability.

Currently we are using the system to develop assignment schedules for the next couple months of 2015 and at the same time we are working on improving and customizing the user interface. While the APS system was created for DAP and using data from that division, it can be easily customized and ported to different pathology departments provided that there is data to revise the model’s parameters. A possible extension of the model would be to consider non-deterministic character of daily requests for pathology services. This extension requires additional data of different granularity and type that, at present, is not easily available at the DAP.

Acknowledgements

This research was supported by the grants from Natural Sciences and Engineering Research Council of Canada and Telfer School of Management Research Fund. The authors would like to acknowledge help and advice of Ms. Joanne Hodgins from the DAP.
References

Appendix

Mixed Integer Linear Programming Model

Decision Variables

\[ X_{ijt} = \begin{cases} 1, & \text{if specimen of type } j \text{ is assigned to pathologist } i \text{ on day } t \\ 0, & \text{otherwise.} \end{cases} \]

\[ i=1,2,3,\ldots,36 \quad j=1,2,3,\ldots,26 \quad t=1,2,3,\ldots,25 \]

\[ Y_{ijk} = \begin{cases} 1, & \text{if pathologist } i \text{ is given specimen of type } j \text{ in week } k \\ 0, & \text{otherwise.} \end{cases} \]

\[ i=1,2,3,\ldots,36 \quad j=1,2,3,\ldots,26 \quad k=1,2,3,4 \]

\( \beta_{ij} \): Auxiliary variable for “Rotation of the specimens” that keeps track of the number of different types of specimens that have been assigned to each pathologist.

Model Parameters

- \( b_i \): the FTE fraction of pathologist \( i \) availability per week
- \( a_j \): the proportion of a full day’s work required for daily pathology requests of type \( j \)
- \( \alpha_{ij} \): equals 1 if specimen type \( j \) belongs to one of the sub-specialties that pathologist \( i \) posses.
- \( T_{it} \): equals 1 if pathologist \( i \) is available on day \( t \), otherwise 0
- \( S_{ij} \): equals 1 if pathologist \( i \) is able to review specimen type \( j \), otherwise 0
- \( C_1, C_2, C_3 \): weights for each part of the objective function

Objective Function

\[
\min z = C_1 \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{K} Y_{ijk} - C_2 \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_{ij} + C_3 \sum_{j=1}^{m} \sum_{t=1}^{T} d_{jt}
\]

The objective function consists of three parts. The first part penalizes assignment schedules each time a pathologist is given a different specimen type in the same week. The second part rewards assignment schedules every time they cover another specimen type that lies within their expertise. The final part penalizes assignment schedules for every uncovered assignment. The weights \( C_1, C_2, C_3 \) allow the user to adjust the importance of the three components to the performance of an assignment schedule. Typically the third component is deemed the most important.
Constraints

1. Consistent assignment in a week (4 weeks within a month, 5 business days within a week):

\[ \forall (i, j) \]
\[ \sum_{t=1}^{5} X_{ijt} \leq MY_{ijt} \]
\[ \sum_{t=6}^{10} X_{ijt} \leq MY_{ij2} \]
\[ \sum_{t=11}^{15} X_{ijt} \leq MY_{ij3} \]
\[ \sum_{t=16}^{20} X_{ijt} \leq MY_{ij4} \]

where \( M \) is a very large number. Each time a new specimen type is assigned to a given week, the above constraints for a new \( Y_{ijk} \) to be positive. Ideally, a pathologist is given only one specimen type in a week so that, for each week, there is only one \( Y_{ijk} \) that is positive.

2. Rotation of the specimens

\[ \forall (i, j) \]
\[ \sum_{t} \alpha_{ij} X_{ijt} \geq \beta_{ij} \]
\[ \beta_{ij} \leq 5 \]

The ideal schedule would give a week’s worth of one specimen type to a pathologist and then switch to a different specimen type for the subsequent week in order to rotate through his/her areas of expertise. This set of constraints rewards schedules that assign a specimen type with a pathologist area of expertise up to a maximum of a reward of 5 for any particular specimen type. This encourages the model to give full weeks of a given specimen and then switch.

3. Weekly assignment

The FTE fractions are per week, but in a scale of 0 to 1, so their values \( (b_i) \) have to be multiplied by 5 in order to derive the number of days that each pathologist works per week (5 business days in a week).

\[ \forall i \]
\[ \sum_{t=1}^{5} \sum_{j=1}^{m} a_{ij} X_{ijt} \leq 5b_i \]
\[ \sum_{t=6}^{10} \sum_{j=1}^{m} a_{ij} X_{ijt} \leq 5b_i \]
\[ \sum_{t=11}^{15} \sum_{j=1}^{m} a_{ij} X_{ijt} \leq 5b_i \]
\[
\sum_{t=16}^{20} \sum_{j=1}^{m} a_j X_{ijt} \leq 5b_i
\]

These constraints force the number of assignments (weighted by the service weight) to respect the number of days that a given pathologist is available based on their FTE fraction.

4. Daily assignments

\[
\forall (i, t) \sum_{j=1}^{m} a_j X_{ijt} \leq 1
\]

This set of constraints ensures that no pathologist is given more than a full day’s work on any given day.

5. Demand coverage

\[
\forall (j, t) \sum_{i=1}^{n} X_{ijt} + d_{jt} = 1
\]

This set of constraints ensures that if a given specimen on a given day (j,t) is not assigned to a particular pathologist then the value of \( d_{jt} \) is forced to be one and thus the objective function captures the number of unassigned slots.

6. Availability

\[
\forall (j, t) \quad X_{ijt} \leq T_{it}
\]

These constraints allow the manager to explicitly state what days a given pathologist may not be available.

7. Skill set

\[
\forall (i, j) \quad X_{ijt} \leq S_{ij}
\]

These constraints again allow the manager to explicitly state which type of specimens a given pathologist is unable to analyse.
Reading and Writing: Qualitative Analysis of Pharmacists' Use of the EHR when Preparing for Team Rounds

Scott D. Nelson, PharmD, MS;1,2 Joanne LaFleur, PharmD, MPH;1,2 Guilherme Del Fiol, MD, PhD;2 R. Scott Evans, MS, PhD, FACMI;2,3 Charlene R. Weir, RN, PhD, FACMI1,2

1Department of Veterans Affairs, Salt Lake City, UT; 2University of Utah, Salt Lake City, UT; 3Intermountain Healthcare, Salt Lake City, UT

Abstract

In the collaborative hospital environment, pharmacists are important members of the healthcare team, yet compared to physicians and nurses, little is known about pharmacists' information needs or how they interact with the electronic health record (EHR). We directly observed seven clinical inpatient pharmacists as they interacted with the EHR preparing for clinical rounds using an eye-tracking camera and contextual inquiry. Pharmacists spent 50% of their time reading information from the EHR, such as notes and medication lists, and 27% of their time copying EHR data onto paper, such as patient history and laboratory results. In an environment where minutes count, the results of this study can help inform the development of CDS tools and/or EHR designs to facilitate the information needs of the pharmacists in providing care for their patients.

Introduction

Pharmacist hospital services, such as medication reconciliation and adverse drug event (ADE) monitoring, are time-intensive processes associated with improved patient care and decreased mortality, making pharmacists integral members of the healthcare team.1-5 Meaningful Use core measures, such as electronic health record (EHR) and computerized provider order entry (CPOE) adoption have impacted the ways pharmacists work and care for patients.6-9 In 2007, only 5.7% of pharmacy departments reported that their institution had a fully electronic patient record;10 this number grew to 26.5% in 2013, with 92.6% reporting having a partial or basic EHR.11 Medication and pharmacy information are critical components of the EHR, and are related to many of the safety features and benefits of EHR use.12 In addition, EHRs have become an important communication tool for interaction between pharmacists, physicians, and nurses.13 Therefore, pharmacists should be included during the development and implementation of EHRs.13-14 Many studies have looked at information needs and seeking behavior of physicians and nurses in the EHR,15-19 but none have evaluated pharmacists. Additionally, the lack of usability or design considerations of health information technology for pharmacists can lead to human-computer interaction problems, increased workflow complexity, loss of productivity, and a decrease in the quality of patient care.15,20

In the Veterans Health Administration (VHA), clinical pharmacists attend clinical patient rounds and participate in patient care planning, which goes beyond medication consultation in many cases. Having a pharmacist present on clinical rounds is a common practice, with over half of all hospitals, and 85% of hospitals with more than 200 beds, having pharmacists attend rounds.11 Pharmacists involvement in clinical rounds has been associated with a significant reduction in ADEs, improved patient care, and reduced costs.3,21-23 In preparation for clinical rounds, pharmacists spend large amounts of time using the EHR to evaluate or “work up” their patients. These evaluations typically consist of identifying potential medication problems, medication regimen review, drug-drug and drug-disease interaction checking, ADE monitoring, therapeutic effectiveness evaluation, dosing appropriateness of medications based on the context of disease states and laboratory values, medication therapy management (MTM), medication reconciliation, and evaluation of patient medication adherence.24 During patient evaluations, pharmacists need to have access to real-time patient information, as other studies have shown that timely information retrieval is important in patient care.25,26 Missing, incomplete, or inaccurate information can lead to medication errors, adverse drug events, failure to provide prophylactic treatment, and other potential patient harm, whereas too much information can cause information overload.15,27 Additionally, distributed and disjointed data can lead to longer times for finding, filtering, and organizing information along with a higher risk of missing important information and cause an increased burden on available resources.16

Due to the abundant and distributed nature of patient data in the EHR, the difficulty of obtaining some data, and time constraints facing pharmacists, an understanding of what information pharmacists need and how they go about finding that information is important. Such an understanding could lead to improved EHR designs or clinical decision support (CDS) tools targeted to pharmacists. Therefore, the purpose of this study is to explore pharmacists’ use of the EHR and information needs in real clinical practice settings using direct observation.
Methods

Study setting and participants
This study was carried out at the George E. Wahlen Department of Veterans Affairs Medical Center in Salt Lake City, Utah. The VHA has a long-established EHR as well as advanced clinical practice for pharmacists, such as independent prescribing privileges within their scope of practice since 1995. The VHA uses the Veterans Health Information Systems and Technology Architecture (Vista) as its EHR, with the Computerized Patient Record System (CPRS) as the clinician interface. A more detailed description of CPRS is available elsewhere. In 2009, the VHA accounted for half of all US hospitals with a comprehensive EHR, and has been associated with improvements in quality of care. Study participants were clinical inpatient pharmacists. We recruited a convenience sample of seven pharmacists from a wide range of inpatient sites such as post-op care, acute medicine, telemetry, rehabilitation, and an intensive care unit. Pharmacists had on average 6.4 years experience with CPRS (range: 1-18, median: 6). Pharmacists were recruited through presentations at staff meetings, demonstrations of the study methods, and referrals from other study participants. We were not able to capture all of the rounds preparation time, but we collected data of the pharmacists evaluating 13 new and 19 familiar patients. We did not collect patient data. This study was approved by the University of Utah Institutional Review Board and VA Research Service.

Procedures
Prior to the observation sessions, pharmacists were given minimal explanation of the study and asked to prepare for clinical rounds as they normally would. Sessions were carried out in the pharmacists' usual environment and time of preparation for rounds, but were limited to approximately the first 45 minutes. We used mixed-methods approach with direct observation in the clinical setting, eye-tracking capture, and contextual inquiry. Direct observation was used because it is difficult to get unbiased responses from providers by directly asking them what their information needs are, or conducting surveys. We encouraged participants to describe their goals or the information they are looking for. However, we did not depend on think-aloud techniques for data collection because in the clinical setting, experts tend to function with system 1 level thought, most people stop talking when they are thinking or processing complex information, and it is difficult for experts to verbalize their goals and tasks. During the session, a researcher would ask clarifying questions to better understand goals, mental models, responsibilities, perceived usefulness of different sources of information, and task. The researcher would wait to ask questions until the pharmacist completed a task, as to not disrupt the cognitive process of complex tasks. In order to capture the rapid, dense, and variable study data, and information sources outside of the EHR, we used an eye-tracking camera from Pupil-Labs to capture video and an iPad to capture audio and additional artifacts. Pupil-Labs is a mobile eye-tracking device with affordable hardware and open source software written in python. The hardware consists of a 3D printed headset that the user wears like a pair of glasses, and has a camera to record the field of view and a camera to record eye movements. The Pupil-Labs camera even allows the researcher to see the pharmacists' gaze positions in real time. The software was run on MacOS, but can be installed on Linux or Windows computers. In addition to the audio and video recording, the researcher would also document field notes during the session and ask deepening questions about information needs and goals after the session. Audio and video from the session were merged and loaded into Atlas.ti 7 (Scientific Software Development GmbH, 2015) for coding and analysis.

Analysis
The analysis of the data was mainly qualitative and descriptive using Atlas.ti for coding and times, and interquartile ranges (IQR) for descriptive statistics. Because of the qualitative and descriptive nature of the study, we estimated that the sample size of seven pharmacists would be sufficient to describe a range of information needs across the inpatient setting. After the observation sessions, the researcher would then code the video using high-level codes to describe what the pharmacist was doing, where they were looking, and what they were writing down. The codes were reviewed and verified by a second researcher. During piloting, we found that the pupil tracker did not provide the precise location of the pharmacist's gaze, but was able to provide a general area of where the pharmacist was looking, which helped in review of the verbal and observational data. This was because the camera calibration suffered when it was taken on and off to visit with patients, and the distance between the tracker and the information source (i.e. computer monitor) changed when the pharmacist moved. Therefore, we did not pursue direct analysis of the eye-tracking data, but rather used gaze data in the context of the rich descriptions in the audio and field notes.

Results
The average time of recorded observation per session was 39 minutes. The average time it took each pharmacist to evaluate a new patient was 11:31 (mm:ss) (median: 10:42, IQR: 9:51 to 14:46). Patients familiar to the pharmacist had an average time of 3:47 (median: 1:46, IQR: 0:56 to 4:51) for review and evaluation in the EHR.
**Reading and writing**

In preparing for rounds, pharmacists spent 50% of their time reading information in the EHR and 27% writing information onto paper, with the rest of their time performing other tasks such as communicating with the team or managing patient lists or notifications. The most predominant information use pattern was the back and forth of reading and writing. Of note, pharmacists switched between reading and writing an average of 4.35 times per minute. In terms of reading, pharmacists spend the most time reading notes in the EHR (such as admission notes, history and physical notes, etc.), followed by medication lists, printed papers (typically print outs of active inpatient and outpatient medication lists), laboratory results (such as reviewing recent results), provider orders (especially over the last 24 hours), and others. However, twice as much time was spent reading EHR notes compared to medication lists. Within the notes, the most time was spent reading the assessment/plan and patient history sections, followed by laboratory results or notes section. The average time spent reading each section was 2.45 minutes, which is significantly higher than the average time spent on medication lists (such as outpatient medications not restarted upon admission), and other medication or historical notes (such as highlighting changes in therapy or identifying preventative treatments). Table 1 shows a list of reading locations and length of time.

<table>
<thead>
<tr>
<th>Reading source</th>
<th>Total time</th>
<th>Time/pharmacist median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notes</td>
<td>43:20</td>
<td>4:34 (4:23-7:55)</td>
</tr>
<tr>
<td>Medication lists</td>
<td>20:32</td>
<td>2:56 (2:05-3:15)</td>
</tr>
<tr>
<td>Paper printouts</td>
<td>16:14</td>
<td>2:48 (2:48-3:04)</td>
</tr>
<tr>
<td>Laboratory results</td>
<td>12:49</td>
<td>1:55 (1:24-2:11)</td>
</tr>
<tr>
<td>Orders</td>
<td>10:28</td>
<td>2:28 (0:23-3:04)</td>
</tr>
<tr>
<td>Vital signs</td>
<td>01:02</td>
<td>0:10 (0:06-0:20)</td>
</tr>
<tr>
<td>Microbiology</td>
<td>00:57</td>
<td>0:14 (0:03-0:25)</td>
</tr>
<tr>
<td>Other sources</td>
<td>05:34</td>
<td></td>
</tr>
</tbody>
</table>

As for writing, significant paper records were kept, including active inpatient medication lists, outpatient medication lists, and medication reconciliation notes. Most information was written down on the active inpatient medication list printed from VistA, not CPRS, and was the main document used on rounds, as shown in Figure 1. Pharmacists spent the most time writing the patient's past medical history (such as problem lists), laboratory results, reminders (such as recommendations or questions for the medical team), medication doses administered (such as frequency of PRN doses, insulin requirements, and IV infusion rates), medication lists (such as outpatient medications not restarted upon admission), and other medication or historical notes (such as highlighting changes in therapy or identifying preventative treatments). Table 2 shows a list of topics written down and times spent writing.

<table>
<thead>
<tr>
<th>Topics written on paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
</tr>
<tr>
<td>Total time</td>
</tr>
<tr>
<td>Time/pharmacist median (IQR)</td>
</tr>
<tr>
<td>Medical history/ problem lists</td>
</tr>
<tr>
<td>Laboratory results</td>
</tr>
<tr>
<td>Reminders</td>
</tr>
<tr>
<td>Medication doses given</td>
</tr>
<tr>
<td>Medication lists</td>
</tr>
<tr>
<td>Medication notes</td>
</tr>
<tr>
<td>Historical notes</td>
</tr>
<tr>
<td>Other topics</td>
</tr>
</tbody>
</table>

**Searching for information**

Search and review strategies varied most in terms of whether the patient was new or familiar. New patients were typically evaluated before familiar patients. When evaluating a new patient, pharmacists typically started in the notes section of the EHR, followed by going to the medication or laboratory results section, followed by the coversheet, orders, or printed medication lists. When evaluating a familiar patient, pharmacists typically started in the orders section to look for updates in treatment, followed by laboratory results or notes section. The average location sequence length was 10 sections of the EHR for new patients, and 5 sections for familiar patients. Additionally, we observed that pharmacists used many different programs in addition to CPRS, such as BCMA, Essentris, VistA, VistA Imaging, and used different non-EHR sources, such as other papers, clinicians, whiteboards, etc. Some information was easier to find than others, for example, in order to find the patient's QTC interval (or even if they had an ECG done) pharmacists had to log into VistA Imaging, which took 51 seconds just to log in.

**Information Tasks**

There were many tasks that the pharmacists performed when preparing for rounds. Some of the most common were comparing medication lists for medication reconciliation, correlating indications/treatments and treatments/indications, managing/optimizing drug therapy (such as insulin or warfarin management), verifying medication administration, monitoring for ADEs, and documenting notes in the EHR (such as medication reconciliation notes). Generally, there was a considerable amount of task switching and multitasking, such as checking for drug-drug interactions, drug doses, medication indications, changes in therapy, etc. when reading the medication list. In this paper, we present managing insulin doses and verifying medication administration as examples of tasks the pharmacists performed in the EHR.
Figure 1. Integrating EHR information by writing notes on an active inpatient medication list printed from VistA. 
A, check list for admission medication reconciliation note and deep vein thrombosis (DVT) prophylaxis. B, reason for admission. C, past medical history or problem list. D, complete blood cell count (CBC) shorthand. E, complete metabolic panel (CMP) shorthand. F, place holders (memory aids) for laboratory results that are pending. G, outpatient medications that have not been restarted. H, triangles in the medication list signal to continued outpatient medications, but with changes compared to outpatient use. I, highlighting DVT prophylaxis. J, circles in the medication list signal to new medications since admission. K, planned duration of antibiotic therapy (some pharmacist would also record which day of regimen). L, "H" in the medication list signals to outpatient medications that were continued on admission. M, items on the list that the pharmacist is not following or going to use on the discharge medication reconciliation note. N, a note that signals to changes in therapy as compared to outpatient.
Example 1. Managing insulin doses

The pharmacists at this institution managed insulin doses, which required an average of 80 seconds to gather needed information to support this task, even though the patients in this study had relatively low insulin requirements. Figure 2 shows a schematic of this task, which required the pharmacist to go to the medication section (or orders section), find one of the insulin orders (such as rapid acting insulin), double click the order, scroll down to the bottom of the popup window to the BCMA data section, and then write down the ancillary blood glucose levels entered as free-text by the nursing staff and write down the insulin doses administered (also free-text entries), as shown in Figure 3. Then they would search for additional insulin orders, such as long-acting or basal insulin, and repeat the above process. The pharmacist would then go to the laboratory results section and write down additional blood glucose levels (which were sometimes different from the results reported in BCMA), and open a laboratory search (or a report) to find the most recent HbA1c lab result, and search for the patient's diet (in the orders or the notes). The pharmacist would then sum the insulin requirements for the last 24 hours, look for trends in the blood glucose values, and provide a recommendation to the team, take no action, or modify the insulin order. To complicate the process, we found that there were multiple insulin orders (such as sliding scale, scheduled, baseline, IV drip, etc.) sometimes using different types of insulin (such as insulin aspart, regular insulin, NPH, or insulin glargine) and that not all blood glucose values were documented in the EHR.

![Example 1. Insulin management process](image)

Figure 2. Example 1. Insulin management process. BG = blood glucose

Example 2. Verifying medication administration.

To verify medication administration, the pharmacists typically had to find the medication order, double click on the order, scroll down to BCMA data and write down times and doses administered. They would check if 'as needed' medications were administered in order to assess pain control or the patient's ability to sleep, check administration times to correlate with drug levels, follow up on medication orders (such as making sure the patient received their warfarin dose), checking when antibiotics were started in order to calculate the treatment duration, and going to the patient's bedside to check IV infusion rates. However, there were some limitations identified. For example, medications administered in the Emergency Department (ED) were not documented in BCMA, so the pharmacists would have to read the ED notes to find doses and administration times. Additionally, BCMA would only pull in the drug concentration and nursing would have to input the dose administered as free text. This was a challenge for one pharmacist when they were trying to find when, or if, the active 80 mg furosemide dose was given to the patient, as shown in Figure 4.
Figure 4. BMCA drug administration data. BCMA data displays drug concentration, but nursing would have to input the dose administered as free text. The pharmacist was searching for 80 mg dose of furosemide.

Discussion

In this study we observed pharmacists as they interacted with the EHR preparing for clinical rounds. We found that pharmacists spent a considerable amount of time searching for and reading information from the EHR and integrating EHR data by writing it down from printed medication lists, suggesting that the way information is stored or displayed in the EHR does not meet the cognitive model or needs of pharmacists.

We observed that not all information in the EHR was useful for pharmacists, or easy to obtain, even when it had been recorded to answer anticipated information needs. For example, CPRS calculates and displays the patient's estimated kidney function using the MDRD creatinine clearance equation. One pharmacist noted that "the MDRD equation is not helpful for me because medication dosing recommendations use the Cockcroft-Gault equation", so he would use an Internet calculator to calculate creatinine clearance using the Cockcroft-Gault equation, which took about 53 seconds. The pharmacists also commented on the not being able to run drug-drug interaction checks on demand when evaluating patient lists. Additionally, the pharmacists typically did not trust the vital signs, fluid balance, or IV infusion rates reported in the EHR, as they were documented infrequently or missing in the EHR.

Other studies of physicians' use of the EHR show that they tend to focus more on the “Assessment and plan” part of clinical notes, and give very little attention to the “Medication profile” part. Ancker et al. also studied physician use of the EHR and found significant physician level variation in utilization of EHR features. Using EHR log data, they found that problem lists were updated in only about 20-25% of encounters, almost always by adding problems instead of removing them. Additionally, problem lists and medication lists were marked as reviewed less than 10% of the time. Given the pharmacists' training and role, it is understandable why pharmacists focus more on medications and medication lists in the EHR. Additionally, pharmacists' focus on medication lists and problem lists to perform daily medication reconciliation and search for drug related problems, such as underdosing or overdosing of medications, ADEs, drug-drug and/or drug-allergy interactions, medication use without an indication, improper medication selection, and untreated medical problems. Similar to physicians, we observed that pharmacists obtained information from a variety of different non-EHR sources, and they typically had multiple systems open at the same time in addition to CPRS. However, having multiple programs open at the same time was problematic as pharmacists tend to jump from one patient to another, and there were instances where the other programs did not update and change patients to be in sync with CPRS, which could have lead to errors.
It was interesting to observe the extensive use of paper and that pharmacists spent about 1/3 of their time transcribing information from the EHR onto paper, especially in an environment that has had a comprehensive EHR in place for over a decade. The persistence of paper is not a new concept, and in 2009, Saleem et al. proposed several reasons for the persistence of paper in the VHA environment, and other benchmark institutions with EHRs. Similar to their study, it seemed that pharmacists mainly used paper because of its ease of use and it helped with efficiency, memory, awareness, task specificity, task complexity, and data organization. For example, most of the pharmacists only had one computer monitor and having information written down on paper helped reduce the need to jump back and forth between tabs in the EHR. The use of paper helped memory by writing things down such as patient history, recommendations, check boxes, medication lists, and placeholders for pending laboratory results. Additionally, paper helped with identifying changes or new information, and allowed for highlighting and quick annotation. Russ et al. suggested that healthcare workers need information to be customizable, prioritized, trendable, locatable, and accessible, requirements that the pharmacists in our study used paper to satisfy.

One study assessed providing pharmacists with tablet computers during rounds, and found many advantages such as the ability to enter/modify/or delete medication orders in real-time, perform complicated calculations, review up-to-date and comprehensive patient data, and access information databases, patient education materials, guidelines, and institutional protocols. However, there were challenges with limited battery life, having to carry around a heavy device, and the limitations of the rigid EHR design, such as searching for information. Even though paper is static, it is cheap, light, flexible, easily carried on team rounds, and not limited by power supply or battery life. The use of paper has important and powerful advantages that cannot be replicated in the EHR and it can help to fill gaps in the EHR design or information; therefore, perhaps we need to better support paper processes using the EHR, such as designing better paper using reports, or bringing related, yet disparate, information in the EHR together to reduce searching/reading time, or tools for associating medications, problem lists, and/or laboratory values. Pharmacists in this study constructed their own custom paper worksheet to help serve as memory aids and facilitate the evaluation process, as shown in Figure 4. In the clinical environment, every minute counts, and tools that can reduce time or increase efficiency are important. The findings in the study, and others, suggest that EHRs provide little cognitive support for clinicians or their workflows.33, 48

![Figure 4. Examples of custom paper worksheets](image)

One limitation of the use of paper by the pharmacists was the creation of handwritten notes and memory aids were more up-to-date than the information in the EHR, and were not available to be handed-off to other pharmacists. When one of the pharmacists was asked about patient handoffs, she said that the other pharmacist will usually call and give a verbal handoff if it is a complex patient, but it is 50:50 if they get a handoff or not, thereby losing the valuable information that the pharmacist spent so much time compiling. Technologies could be developed, such as...
EHRs or EHR apps, that are easily developed and customized, allowing clinicians to create dashboards or reports that are relevant to their setting or role, thereby becoming part of the information system instead of persisting alongside it.\textsuperscript{50} Some have reasonably suggested that it is not possible to meet the information needs of all users, as different users have different workflows;\textsuperscript{51} however, making the EHR customizable for different use cases, such as through apps or CDS tools that access core services and data in the EHR, could make it possible to meet the information and workflow needs of a spectrum of various users.\textsuperscript{52, 53}

Lastly, Millonig\textit{ et al.} suggested that pharmacists need to have access to patient-specific health information to improve patient outcomes, including a medication profile with refill history, vital signs and laboratory test results, diagnoses and current conditions, and medical history and physical assessment data.\textsuperscript{39} While our project focused on the inpatient setting, the availability of patient-specific information to help satisfy pharmacist information needs in the ambulatory and community settings could help improve patient outcomes, especially since pharmacists are more likely to detect and prevent errors by having more access to patient information that provides the overall context of the patient.\textsuperscript{54, 55} Pharmacists, and other clinicians, should have access to the entire patient record, regardless of where care was received, such as potential outpatient medication lists and problem lists.\textsuperscript{5} As pharmacists’ scope of practice continues to expand, especially in other institutions, so too does their need for adequate patient information as meaningful users and contributors to the EHR.\textsuperscript{54, 56-58}

\textbf{Limitations}\n
There are some limitations to this study that should be mentioned. Frist, we had a small sample size from only one institution, and we were not able to explore site-specific cultural needs or additional individual user preferences. Additionally, the generalizability of our results to non-VHA institutions is limited due to the unique characteristics of CPRS and the exemplary example of progressive clinical pharmacy practice in the VHA. The practice of pharmacy, roles of pharmacists, and EHR data collected or reported may be different at other institutions. However, anecdotal experience at local surrounding non-VHA hospitals suggests similar information needs and use of paper. Additionally, we were not able to have independent coding by the second researcher. The first researcher had significantly more knowledge and context of the sessions because he was a pharmacist and he was the one present for the observations; however, both coders had to agree on consensus for the coding procedures. While this project focused on the actual user workflow, future research could compare these results with the pharmacists’ perception of an ideal workflow in their clinical area of focus.

\textbf{Conclusion}\n
Pharmacists spend a significant amount of time searching for patient information in the EHR, and writing it down onto paper. This study suggests that current EHR designs do not readily support the information needs or workflow requirements of pharmacists. In an environment where every minute counts, the results of this study can help inform the development of CDS tools and/or EHR designs to facilitate the information needs of the pharmacists in providing care for their patients.

\textbf{Acknowledgement}\n
During this study, Dr. Nelson was supported by the VA Advanced Fellowship Program in Medical Informatics of the Office of Academic Affiliations, Department of Veterans Affairs.

\textbf{Disclaimer}\n
The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

\textbf{References}\n

950


13. Hagland M. Is there a pharmacist house? When it comes to clinical IT implementations, healthcare IT leaders are turning to the 'third discipline' for leadership and involvement in clinical informatics. Healthc Inform. 2010;27(12):18, 20, 2, passim.


36. Kassner MP, Patera WR. PUPIL: constructing the space of visual attention: Massachusetts Institute of Technology; 2012.
Assessing the Utility of Automatic Cancer Registry Notifications Data Extraction from Free-Text Pathology Reports

Anthony N. Nguyen, PhD1, Julie Moore2, John O'Dwyer1, Shoni Philpot2
1The Australian e-Health Research Centre, CSIRO, Brisbane, Australia
2Queensland Cancer Control Analysis Team, Queensland Health, Brisbane, Australia

Abstract
Cancer Registries record cancer data by reading and interpreting pathology cancer specimen reports. For some Registries this can be a manual process, which is labour and time intensive and subject to errors. A system for automatic extraction of cancer data from HL7 electronic free-text pathology reports has been proposed to improve the workflow efficiency of the Cancer Registry. The system is currently processing an incoming trickle feed of HL7 electronic pathology reports from across the state of Queensland in Australia to produce an electronic cancer notification. Natural language processing and symbolic reasoning using SNOMED CT were adopted in the system; Queensland Cancer Registry business rules were also incorporated. A set of 220 unseen pathology reports selected from patients with a range of cancers was used to evaluate the performance of the system. The system achieved overall recall of 0.78, precision of 0.83 and F-measure of 0.80 over seven categories, namely, basis of diagnosis (3 classes), primary site (66 classes), laterality (5 classes), histological type (94 classes), histological grade (7 classes), metastasis site (19 classes) and metastatic status (2 classes). These results are encouraging given the large cross-section of cancers. The system allows for the provision of clinical coding support as well as indicative statistics on the current state of cancer, which is not otherwise available.

Introduction
Cancer notified from pathology is the primary method of identifying population based cancer incidence and is an important and fundamental tool for cancer monitoring, service planning and research. The Cancer Registry receives cancer specimen reports from pathology laboratories, which are subsequently abstracted by expert clinical coders for key cancer characteristics. The information is often trapped in the language of these reports, which are in the form of unstructured, ungrammatical and often fragmented free-text. The effort required for information abstraction can therefore be an extremely labour and time intensive exercise. Furthermore, the abstraction is also subject to errors and inconsistent interpretations due to the need for repeated interpretation of the results by coders with differing levels of experience and training potentially leading to differing conclusions, repeated data entry into collection systems, and when cases are misinterpreted or keywords are missed.

An approach whereby reports are electronically received and automatically processed, abstracted and analysed has the potential to support expert clinical coders in their decision-making and assist with improving accuracy in data recording. Improving the cancer notifications process would provide significant benefits to oncology service providers, health administrators, clinicians and patients.

An automated medical text analysis system that extracts cancer notifications data from any notifiable electronic cancer pathology report is proposed. A rule-based approach utilising natural language processing (NLP) and symbolic reasoning using SNOMED CT were adopted in the system. Selected Queensland Cancer Registry business rules were also incorporated to mimic the interpretations and coding standards that expert clinical coders would adopt. The system was deployed to process pathology HL7 feeds from across the state of Queensland in Australia. The utility of the system was assessed and showed promising results on a set of reports containing a large cross-section of cancers.

Background
There has been a number of clinical language processing systems or studies relating to the extraction of key cancer characteristics from pathology free-text. Most research has focused on data extraction tasks for specific cancers such as colorectal, breast, prostate and lung.

* Systematized nomenclature of medicine - clinical terms
The medical text analysis system/pathology (MedTAS/P) proposed by Coden et al. uses NLP, machine learning and rules to automatically extract or classify cancer characteristics. Selected cancer characteristics were evaluated and showed promise with F-measures ranging from 0.9–1.0 for most extraction tasks including histological type, primary site, and grade on a corpus of colon cancer pathology reports.

Martinez and Li, similarly, used a colorectal cancer database to automatically predict cancer characteristics using machine learning (and in some cases complemented with rules) with 5 of the 6 multiclass problems achieving an F-measure above 74.9% using simple feature representations. Primary site, however, proved difficult to predict with an F-measure of 0.58.

Ou and Patrick extracted pertinent colorectal cancer information from narrative pathology reports using supervised machine learning and automatically populated the cancer structured reporting template using rule-based methods. They achieved an overall F-measure of 81.84% over a large range of structured reporting data fields.

Currie et al. presented a method of automated text extraction using specific rules and language patterns to extract over 80 data fields from breast and prostate cancer pathology reports with 90-95% accuracy for most fields.

Buckley et al. studied the feasibility of using natural language processing to extract clinical information from over 76,000 breast pathology reports from 3 institutions. They reported that there was widespread variation in how pathologists reported common pathologic diagnoses. For example, 124 ways of saying ‘invasive ductal carcinoma’, 95 ways of saying ‘invasive lobular carcinoma’ and over 4000 ways of saying ‘invasive ductal carcinoma was not present’. Reported sensitivity and specificity of the system were 99.1% and 96.5% when compared to expert human coders.

The Medical Text Extraction (Medtex) pipeline proposed by Nguyen et al. used a symbolic rule-based approach to parse pathology reports using NLP to identify SNOMED CT concepts of relevance, and tested whether these concepts were subsumed by concepts relating to cancer staging factors. Lung cancer staging and synoptic reporting were used to illustrate the symbolic rule-based approach. The symbolic rule-based system performed within the bounds of human staging accuracy as observed in studies of registry data.

As these studies are cancer (or tumour stream) specific, more generalized approaches are needed to extract cancer characteristics for all cancers. More recent research using Medtex has been applied to the extraction and coding of cancer characteristics such as basis of diagnosis, histological type and grade, cancer site and laterality from pathology free-text for all cancers. Preliminary results on a small evaluation set of 61 cancer notifiable reports comprised of a range of cancers have shown that cancer characteristics can be extracted with an overall accuracy of 80%.

In this paper, we present the architecture and deployment of Medtex on streaming pathology HL7 feeds from public pathology laboratories across the state of Queensland, Australia. Challenges here included the vast individual pathologists and institutional variations in the textual contents of the reports. A subset of 220 pathology reports from the deployment was selected from patients with a range of cancers to evaluate and analyse the performance of the system over seven cancer characteristic categories, each potentially containing a large number of possible classes, namely, basis of diagnosis (4 classes), primary site (330 classes), laterality (5 classes), histological type (1036 classes), histological grade (9 classes), metastasis site (330 classes) and metastatic status (2 classes). In contrast to previous work, the work presented here evaluated the utility of a Medtex deployment using a different and larger evaluation dataset and, unlike all previous tumour stream specific studies, this paper presents cancer characteristic extraction results on a wide range of cancer sites and types. The robustness of the system is also presented by comparing the evaluation results against those obtained from the development set and from a majority class classifier. An error analysis of the poorer performing cancer characteristic categories was also performed to determine the underlying limitations of the system.

Method

System Description

The medical text analysis system, Medtex, is a Java-based NLP software platform created for the development of clinical language engineering analysis engines to support data-driven analytic tasks. Medtex incorporates a (1) free-text to SNOMED CT mapping engine to normalize the free text (i.e. unify the language of the reports) by identifying medical concepts, abbreviations and acronyms, short-hand terms, dimensions and relevant legacy codes, (2) relate key medical concepts, terms and codes using contextual information and report substructure, and (3) use formal semantics, via a SNOMED CT ontology server, for medical text inference and reasoning. Additional analysis
engines can be incorporated to infer or classify complex clinical notions relevant to a particular health application using handcrafted algorithms and rules and/or machine learning techniques. Medtex has been applied to small scale datasets for research purposes; however its utility on real-time data streams and larger datasets may be inadequate if the computational time for the analysis of reports cannot keep up with the demands of the incoming data stream.

To address this issue, the Java messaging service (JMS) was chosen as the messaging broker for providing an intermediary to allow Java applications to be loosely coupled and reliably create, send, receive and read messages. This messaging service is built on the concept of message queues, producers (senders), and consumers (receivers). A message producer is used for sending messages to a specific queue. The message consumer is then used for receiving messages from a specified queue. Multiple message consumers can be set up in parallel to receive messages from the same queue such that only one message is received by only one of the consumers. Furthermore, consumers acting on data can publish their results to another queue called a message topic, whereby other consumers wishing to register and subscribe to the topic can receive messages from the topic. This scenario allows for multiple consumer applications to act on the same messages published from a given consumer.

The proposed Medtex service for analysing HL7 messages from a statewide pathology information system is illustrated in Figure 1. It aims to automate a number of Cancer Registry tasks such as the notification of cancer reports and the coding of notifications data. Apache ActiveMQ, an open source message broker, which fully implements JMS, was used to implement the messaging service. The message producer (HL7 Producer) accesses pathology HL7 messages and through the selection of report types that are relevant for subsequent processing, messages are sent to a specified queue (REPORTS_QUEUE). Multiple Medtex consumers can be set-up such that each consumer will take a message from the queue in turn. The results from the Medtex analysis are encoded in JSON format and published to a message topic (RESULTS_SUBSCRIPTION) where the topics can be subscribed to by end-user applications (Results Consumer), for example, to consolidate patient results and store them in a database and/or provide support for clinical coders to abstract clinical information from medical reports.

Within Medtex itself, the system would automatically process and analyse free-text HL7 pathology reports. The system selects from the pathology feed, histology and cytology reports, and filters non-notifiable cancer reports. For notifiable cancer reports, cancer characteristics extracted for a Cancer Registry notification consist of basis of diagnosis, primary site, laterality, histological type, histological grade, metastasis site, metastatic status, among others. Those relevant for the evaluation in the current study include:

- **Basis of diagnosis** encodes the method by which the cancer was diagnosed. For cytology and histology reports, the basis of diagnosis can either be encoded as cytology or haematology (06), histology of metastasis (07), histology of primary (08), or autopsy and histology (09). The basis of diagnosis is often used to assess the reliability of the cancer diagnosis, where the most conclusive information is found from histological reports.
- **Histological type** records the characteristics of the tumour. It is encoded using an ICD-O morphology code, which consist of a prefix M followed by a five-digit code ranging from M-80000 to M-99893. The cell type and behaviour make up the morphology code. The first 4 digits refer to the histology, while the last digit is

---

‡ [http://www.json.org/](http://www.json.org/)
the behaviour code and identifies whether the neoplasm is benign (0), uncertain and unknown behaviour (1), in situ (2), malignant (3), secondary or metastatic (6), or malignant but unknown whether it’s a primary or metastatic site (9). Behaviour codes 6 and 9 are not used by Cancer Registries but are instead flagged or derived from other cancer characteristic data elements such as basis of diagnosis, metastatic site or metastatic status.

- **Histological grade**, differentiation or phenotype describes how much or how little a tumour resembles normal tissue. It is encoded using a one-digit code. Only malignant tumours are graded and are represented by code numbers 1 to 4, designating grades I (well differentiated) to IV (undifferentiated), respectively. For a lymphoma or leukaemia, separate code numbers 5 to 8 are used to identify immunophenotype differentiation such as T-cell, B-cell, Null cell, and NK cell origin, respectively. If grading is unknown, not applicable or cannot be determined, then a code number of 9 would be assigned.

- **Primary site** describes the origin of the cancer in the body and is represented by an ICD-O topology code ranging from C00.0 to C80.9. It is encoded using a four-character code using a prefix C to identify topography codes. The first two digits represent the site (e.g. C34 for Lung), while the last digit defines the sub-site (e.g. C34.1 for Upper lobe of lung).

- **Laterality** indicates the side affected by the tumour for cancers of paired organs (e.g. breast, lung, kidney, etc.). It is encoded using a one-digit code: right (1), left (2), bilateral (3), not applicable (8) and unknown (9).

- **Metastasis site** describes the site of spread from which the cancer originated. In this study, it was proposed to assign an equivalent ICD-O topography code to represent the metastatic site, although ICD-O is not usually used for this purpose.

- **Metastatic status** is a flag to reflect whether the behaviour of a tumour indicates a metastasis (including lymph node metastasis).

The cancer characteristic codes were defined from ICD-O Third Edition for primary site, histological type and histological grade; and other notifications data according to classification codes recorded in the Queensland Cancer Registry. Table 1 summarises the list of the cancer characteristics along with their codes (or classes). In general, the cancer characteristic categories are multiclass where there are more than 2 classes within the category. The vast number of histological types and primary sites for classification show the complexity of the extraction tasks involved.

<table>
<thead>
<tr>
<th>Category</th>
<th>Code</th>
<th>Number of Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basis of Diagnosis</td>
<td>06 – Cytology or Haematology</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>07 – Histology of metastasis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>08 – Histology of primary</td>
<td></td>
</tr>
<tr>
<td></td>
<td>09 – Autopsy and histology</td>
<td></td>
</tr>
<tr>
<td>Histological Type</td>
<td>ICD-O morphology code – M-xxxxx</td>
<td>1036</td>
</tr>
<tr>
<td>Histological Grade</td>
<td>1 – Grade I – well differentiated</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2 – Grade II – moderately differentiated</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 – Grade III – poorly differentiated</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 – Grade IV – undifferentiated or anaplastic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 – T-cell</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 – B-cell, Pre-B, B-precursor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 – Null cell, Non T, Non-B (For leukaemias only)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 – NK Cell</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9 – Grade or differentiation not determined, not stated or not applicable.</td>
<td></td>
</tr>
<tr>
<td>Primary Site</td>
<td>ICD-O topography code – Cxx.x</td>
<td>330</td>
</tr>
<tr>
<td>Laterality</td>
<td>1 – Right</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2 – Left</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 – Bilateral</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 – Not applicable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9 – Unknown</td>
<td></td>
</tr>
<tr>
<td>Metastatic Site</td>
<td>Equivalent ICD-O topology code – Cxx.x</td>
<td>See primary site</td>
</tr>
<tr>
<td>Metastatic Status</td>
<td>Not applicable or 2 (metastasis)</td>
<td>2</td>
</tr>
</tbody>
</table>

An expert clinical coder analysed the development set (see Corpus Description) to help build the ground truth and extraction modules for each of the cancer characteristic categories. A combination of NLP, domain knowledge and rules, and in particular SNOMED CT manipulation and querying were used to classify cancer characteristics. The
algorithm and rules were iteratively refined based on measuring and analysing the performance of the system on the development data set. Examples of the cancer characteristic classification sites are tabulated in Table 2.

Table 2. Example of methods used for the extraction of cancer notifications data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Notifications Data</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queensland Cancer Registry coding rules (including special casings)</td>
<td>Histological Type</td>
<td>Select the highest morphology if more than one morphology is stated.</td>
</tr>
<tr>
<td></td>
<td>Histological Grade Primary Site</td>
<td>Assign the highest grade or differentiation code.</td>
</tr>
<tr>
<td>Domain knowledge</td>
<td>Primary Site</td>
<td>Code all leukaemia except myeloid sarcoma (M-99303) to C42.1 (bone marrow).</td>
</tr>
<tr>
<td>SNOMED CT property access</td>
<td>Histological Type</td>
<td>List of one-to-one only ICD-O morphology to site mappings.</td>
</tr>
<tr>
<td></td>
<td>Primary Site</td>
<td>Restrict SNOMED CT concepts to those with a ‘morphologic abnormality’ semantic category and those that have alternate terms with the following regular expression “M{[0-9][5]}”.</td>
</tr>
<tr>
<td>SNOMED CT to ICD-O topography cross-maps</td>
<td>Primary Site</td>
<td>Map SNOMED CT ‘body structure’ concepts to ICD-O topography codes.</td>
</tr>
<tr>
<td>SNOMED CT Subsumption querying</td>
<td>Histological Type</td>
<td>Candidate ‘leukaemia’ concepts are found by testing subsumption by the ‘128931003</td>
</tr>
<tr>
<td>SNOMED CT Concept relationship querying</td>
<td>Primary Site</td>
<td>“Procedure site – Direct” and “Finding site” relationship values from concepts are used as candidate sites.</td>
</tr>
<tr>
<td>SNOMED CT querying using ad-hoc term expansion</td>
<td>Histological Type</td>
<td>The histological type and grade’s preferred terms were used to search for a more specific concept. For example, the query for “Follicular Lymphoma” + “Grade 3” would return the histological type M-96983, which is “Follicular Lymphoma, Grade 3”.</td>
</tr>
<tr>
<td>Relation extraction</td>
<td>Basis of Diagnosis</td>
<td>Identification of multiple concepts or terms within a search scope such as metastasis and lymph nodes within a sentence.</td>
</tr>
<tr>
<td>Keyword/phrase spotting</td>
<td>Histological Grade</td>
<td>Detect keywords or phrases that were unable to be (or unreliably) mapped to SNOMED CT. For example “poorly to moderately differentiated”.</td>
</tr>
</tbody>
</table>

Corpus Description

Access to the Queensland statewide pathology data was obtained from the Queensland Oncology Repository with research ethics approval from the Queensland Health Research Ethics Committee. The data covers HL7 pathology feeds from public pathology laboratories in the state of Queensland. A corpus consisting of 500 pathology reports was used for system development of which 201 of them were notifiable cancers (and thus relevant for the current cancer characteristic extraction task). Non-notifiable cancers such as non-malignant cancers, and squamous cell carcinoma (SCC) and basal cell carcinoma’s (BCC) of the skin were identified, but removed and flagged by the system; a separate study addressed these issues by filtering notifiable reports from non-notifiable reports. For system evaluation, a separate 220 pathology reports from the deployment of the system for processing the backlog of pathology feeds was selected from patients with a range of cancers (i.e. tumour-stream stratified sampling) to evaluate and analyse the performance of the system. The ground truth used for system evaluation was based on the reference data set annotated by the same expert clinical coder who helped develop the system. Table 3 shows the cancer characteristic statistics from the development and evaluation corpus.

Table 3. Notifiable cancer characteristic corpus statistics.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Classes</th>
<th>Majority Class</th>
<th>Frequency Range (Mean ± Std Dev)</th>
<th>Number of Unseen Classes in Eval.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basis of Diagnosis</td>
<td>3</td>
<td>08</td>
<td>19-146 (67±69)</td>
<td>21-175 (73±88)</td>
</tr>
<tr>
<td>Histological Type</td>
<td>64/94</td>
<td>M-81403</td>
<td>1-3 (3.1±5.4)</td>
<td>1-21 (2.3±3.1)</td>
</tr>
<tr>
<td>Histological Grade</td>
<td>7</td>
<td>9</td>
<td>3-110 (28.7±38.1)</td>
<td>1-129 (31.4±44.6)</td>
</tr>
<tr>
<td>Primary Site</td>
<td>58/66</td>
<td>C50.9</td>
<td>1-21 (3.4±4.1)</td>
<td>1-39 (3.3±5.9)</td>
</tr>
<tr>
<td>Laterality</td>
<td>4/5</td>
<td>8</td>
<td>20-129 (50.3±52.6)</td>
<td>1-134 (44.0±52.4)</td>
</tr>
<tr>
<td>Metastatic Site</td>
<td>17/19</td>
<td>NA</td>
<td>1-170 (11.2±39.7)</td>
<td>1-192 (11.0±42.6)</td>
</tr>
<tr>
<td>Metastatic Status</td>
<td>2</td>
<td>NA</td>
<td>31-170 (101±98)</td>
<td>28-192 (110±116)</td>
</tr>
</tbody>
</table>

Dev., development set (N=201); Eval., evaluation set (N=220); Std Dev, standard deviation; NA, ‘Not Applicable’ class

The distributions between the development and evaluation set for each cancer characteristic category were quite similar. However, within the categories the class frequencies can vary quite significantly, with large variation in ranges and standard deviations, e.g. basis of diagnosis, histological grade, laterality, etc. On the other hand, in some categories, there are a large number of small frequency classes resulting in low means and standard deviations, e.g. histological type and primary site, due to the large and diverse number of cancer types and sites, respectively.
Furthermore, there is a large portion of classes contained in the evaluation set that were not found in the development set (e.g. histological type, primary site and metastatic site). The corpus and cancer characteristic category statistics show the challenges and complexity of the extraction tasks. In addition, the varied writing style and language contained within the reports from different pathologists and laboratories creates additional challenges that will test the robustness and generalizability of the extraction modules when measuring the performance of the system on the evaluation set.

**Performance Measures**

The measures used to evaluate results are based on the counts of true positives ($TP$), true negatives ($TN$), false positives ($FP$), and false negatives ($FN$) resulting from the classification decisions. The multiclass classification performance for a given cancer characteristic category, $C$, is measured using the micro-average recall ($R_{micro}$) or sensitivity), precision ($P_{micro}$ or positive predictive value), and balanced F-measure ($F$).

$$R_{micro} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FN_i}; \quad P_{micro} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i}; \quad F = \frac{2PR}{P + R}$$

Overall system performance is reported as the macro-average measure across all the categories, with equal weight to every category.

A majority class classifier was used as a baseline to determine whether the system’s results were significantly better than one that naively classifies results simply based on the majority class. In addition, to assess the generalizability of the system, the system’s results from the evaluation set were compared with that from the development set. Other cancer notification extraction systems are tumour specific and thus were not suitable for use as a benchmark.

**Results**

The Medtex messaging service was applied to a statewide pathology HL7 message feed. The backlog and new incoming HL7 messages from pathology laboratories in Queensland, Australia were used to analyse the pathology reports as well as test the load on the service. Using 3 instances of Medtex, the system’s average processing rate was 3.6 seconds per message and achieved the processing of a year’s worth of messages within just under 5 days.

An increase in report analysis throughput was achieved by using the messaging framework and multiple instances of Medtex consumers in parallel. The use of 3 Medtex instances in parallel resulted in a 2.5 times speed-up over the sequential single instance of Medtex in operation. Depending on system resources, further speed-ups are possible if additional instances of Medtex and/or multiple instances of Medtex’s shared resources such as the SNOMED CT ontology and concept-mapping servers were made available.

The classification performances of the system with respect to the cancer characteristic categories are shown in Table 4. Figure 2 summarises the F-measure comparison of cancer characteristic categories between the development and evaluation set.

<table>
<thead>
<tr>
<th>Cancer Characteristic</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basis of Diagnosis</td>
<td>0.955</td>
<td>0.918</td>
<td>0.965</td>
</tr>
<tr>
<td>Histological Type</td>
<td>0.796</td>
<td>0.577</td>
<td>0.865</td>
</tr>
<tr>
<td>Histological Grade</td>
<td>0.935</td>
<td>0.773</td>
<td>0.945</td>
</tr>
<tr>
<td>Primary Site</td>
<td>0.522</td>
<td>0.546</td>
<td>0.656</td>
</tr>
<tr>
<td>Laterality</td>
<td>0.786</td>
<td>0.805</td>
<td>0.794</td>
</tr>
<tr>
<td>Metastatic Site</td>
<td>0.891</td>
<td>0.886</td>
<td>0.918</td>
</tr>
<tr>
<td>Metastatic Status</td>
<td>0.945</td>
<td>0.932</td>
<td>0.945</td>
</tr>
<tr>
<td>Macro-average</td>
<td>0.833</td>
<td>0.777</td>
<td>0.870</td>
</tr>
</tbody>
</table>
The system’s results were compared to a majority class classifier to determine whether the system was significantly better than one that naively classifies results simply based on the majority class. Table 5 shows the F-measure results for both the system and the majority class classifier.

Table 5. Comparison of F-measure results between the system (Medtex) and majority class classifier.

<table>
<thead>
<tr>
<th></th>
<th>Development (N=201)</th>
<th>Evaluation (N=220)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Majority Classifier</td>
<td>Medtex</td>
</tr>
<tr>
<td>Basis of Diagnosis</td>
<td>0.73 (0.66-0.79)</td>
<td>0.96 (0.93-0.99)</td>
</tr>
<tr>
<td>Histological Type</td>
<td>0.17 (0.12-0.23)</td>
<td>0.83 (0.78-0.88)</td>
</tr>
<tr>
<td>Histological Grade</td>
<td>0.55 (0.48-0.62)</td>
<td>0.94 (0.91-0.97)</td>
</tr>
<tr>
<td>Primary Site</td>
<td>0.11 (0.06-0.15)</td>
<td>0.58 (0.51-0.65)</td>
</tr>
<tr>
<td>Laterality</td>
<td>0.64 (0.58-0.71)</td>
<td>0.79 (0.73-0.85)</td>
</tr>
<tr>
<td>Metastatic Site</td>
<td>0.85 (0.80-0.90)</td>
<td>0.90 (0.86-0.95)</td>
</tr>
<tr>
<td>Metastatic Status</td>
<td>0.85 (0.80-0.90)</td>
<td>0.95 (0.91-0.98)</td>
</tr>
<tr>
<td>Macro-average</td>
<td>0.56 (0.49-0.62)</td>
<td>0.85 (0.80-0.90)</td>
</tr>
</tbody>
</table>

Discussion

Overall system performance on the evaluation set reported as F-measure was 0.80. At a cancer characteristic level, the F-measure performances within each category were 0.93 for basis of diagnosis, 0.64 for histological type, 0.79 for histological grade, 0.61 for primary site, 0.82 for laterality, 0.90 for metastatic site and 0.93 for metastatic status. The results are promising given the challenges previously discussed regarding the large number of classes from certain categories, varied and skewed distributions within each cancer characteristic category, and the large number of unseen classes being classified in the evaluation set. The results show the system’s robustness and generalizability by achieving extraction performances on the evaluation set that is comparable with that obtained from fine-tuning the system using the development set, and also its superiority to that obtained when using a majority class classifier. When compared to previous studies that focused on certain tumour streams, the results show that generalizing the extraction algorithms to accommodate for all tumour streams has its challenges and therefore is sub-optimal to the tumour stream specific results. However, it would be a very costly exercise to build specific tumour stream cancer characteristic classifiers for each and every possible cancer. The trend in extraction performances across the categories are also consistent with previous works where primary site was found to be the most challenging.

The results between the development and evaluation set were in general not significantly different. Only the histological type and histological grade generated results that exhibited non-overlapping 95% confidence intervals suggesting that the algorithms and rules adopted in these extraction modules likely over-fitted the development data. That said, there was a large number of histological types in the evaluation set that were not seen in the development set; despite having almost 70% of the histological types unseen by the system during development, the system was able to classify the category with a recall of 0.58, precision of 0.71 and an overall F-measure of 0.64. In terms of the
free-text to SNOMED CT mapping engine, it was observed that 83.2% of the histological types from the
development set could have been found within the mapped concepts. This provides an upper bound to the
classification performance by the system, unless other methods are introduced to infer the histological type.

For histological grade, error analysis revealed that at a per-class level, all classes had non-overlapping 95% confidence intervals, except for the histological grade of 3 (poorly differentiated). The errors were more pronounced for histological grades 4 through to 6, from which 5 and 6 relate to lymphoma or leukaemia histological types.

The extraction of primary site was also a challenge for the system as evident from its performance. Despite having 45% of the cases in the evaluation set unseen by the system during development, the results between the development and evaluation set had overlapping 95% confidence intervals. One source of error was due to the lack of co-referencing of specimens between the macroscopic and microscopic sections of the pathology report. As a result, the relation between different evidences in the free-text could not be classified correctly. Again in terms of the free-text to SNOMED CT mapping engine, 70.3% of the primary sites (excluding unknown primary C80.9 sites) from the development set were found in the mapped concepts. This suggests that the use of the current concept-mapping algorithm has its limitations in giving the system the ability to correctly identify primary sites. Further inspection of the errors also revealed that many of the cases actually classified the ICD site = C34 for Lung) but not the sub-site (Cxx.x; e.g. sub-site = C34.1 for Upper lobe of lung). Figure 3 illustrates this effect where results at a site level (Cxx) were significantly better than that classified at a sub-site level (Cxx.x) with non-overlapping 95% confidence intervals.

![Figure 3. Primary site classification performances at a site (Cxx) and sub-site (Cxx.x) level along with 95% confidence intervals.](image)

The metastatic site and metastatic status data items are also worth deeper analysis. Although, performing well with high F-measures, the large bias towards the majority class did not make the system perform much better than the majority class classifier. As a result, it is likely that there is poor recall and precision for the minority classes.

Future research will need to focus on iteratively improving the system’s performance, especially for critical cancer notifications data such as primary site and histological type. Preliminary error analysis have revealed areas of focus for further system improvement such as histological grade for lymphoma and leukaemia cancers, including the need to investigate other concept-mapping algorithms for improving both histological type and primary site categories. These error analyses facilitate the identification of the relevant Queensland Cancer Registry business rules to be incorporated, abstraction errors by human experts, and also feedback the type of errors generated for further system development.

Other limitations of the system include limited number of development reports given the large number of possible histological types and primary sites. The evaluation corpus is also biased towards uncommon cancers due to the tumour-stream stratified sampling that was applied to ensure that a range of cancers were represented in the evaluation set. In addition, the system at present cannot distinguish between multiple cancers reported within a single report.

In a clinical coding workflow setting, the system could be used to support clinical coders and hence improve data collection capture at Cancer Registries by highlighting and pre-populating cancer notification items for validation (for example, see Figure 4). Here, the free-text report is shown in the leftmost panel. The highlights shown over the
free-text report correspond to the evidence used to generate the system’s suggested coding, which is shown in the rightmost panel. Clinical coders can then use the system’s suggestions to either accept (using the arrows in the rightmost panel) or enter in an alternate code to populate the fields within the clinical coder’s coding panel (as depicted by the middle panel). Investigations into the utility of a guided and interactive annotation process on cancer characteristic abstraction tasks is reserved for future work.

Figure 4. The Medtex software can process narrative pathology reports and generates structured data to aid clinical coders in cancer abstraction tasks.

The proposed architecture for the processing and analysis of streaming pathology reports has potential to overcome the multi-year delay in the reporting of cancers by providing indicative population-level statistics on the current incidence of cancer. The system supports future data extension requirements such as cancer stage, and also can be applied on other sources of cancer data with free text, e.g. death data and radiology reports.

Conclusion

Analysis of the contents of electronic pathology reports will have a profound impact on cancer care. As a result, a number of automatic cancer data extraction systems have been developed, however their utility in automating Cancer Registry tasks has to be adequately assessed. In this study, the Medtex system was assessed and showed promise in terms of stream processing at a statewide level and also in terms of cancer characteristic extraction performances and its ability to track and identify specific system limitations for future improvements.

The use of Medtex’s messaging technologies that take advantage of the parallelism of consumers/producers can be an effective real-time processing solution for data streams. These technologies greatly increase the throughput of medical text analytics for clinical decision support and/or research activities involving real-time data streams or large datasets.

The cancer notifications data extraction results from Medtex show promise with an overall F-measure performance of 0.80 on a broad range of cancers and cancer characteristic categories. Despite some cancer characteristic categories performing well, with an F-measure score of above 0.90, the histological type and primary site extraction results proved more challenging due to its large number of possible classes. The system is extensible and cancer stage including other synoptic data can also be incorporated. Future work will analyse errors between the system and the reference standard to feedback into the iterative development process.

The system is proposed to streamline and support the clinical coding workflow at Cancer Registries by identifying cancer notifiable reports and then highlighting and pre-populating cancer notification items for clinical coder validation. It is hoped that such automation would help overcome the multi-year delay in the reporting of cancer
statistics with Cancer Registries able to have access to up-to-date population-level statistics on the current state of cancer.

References

Nutrition Informatics Applications in Clinical Practice: a Systematic Review

Jennifer C. North MS, RD, CD¹, Kristine C. Jordan PhD, MPH, RD¹, Julie Metos PhD, MPH, RD¹, John F. Hurdle MD, PhD²

¹Division of Nutrition, University of Utah, Salt Lake City, Utah, United States; ²Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, United States

Abstract

Nutrition care and metabolic control contribute to clinical patient outcomes. Biomedical informatics applications represent a way to potentially improve quality and efficiency of nutrition management. We performed a systematic literature review to identify clinical decision support and computerized provider order entry systems used to manage nutrition care. Online research databases were searched using a specific set of keywords. Additionally, bibliographies were referenced for supplemental citations. Four independent reviewers selected sixteen studies out of 364 for review. These papers described adult and neonatal nutrition support applications, blood glucose management applications, and other nutrition applications. Overall, results indicated that computerized interventions could contribute to improved patient outcomes and provider performance. Specifically, computer systems in the clinical setting improved nutrient delivery, rates of malnutrition, weight loss, blood glucose values, clinician efficiency, and error rates. In conclusion, further investigation of informatics applications on nutritional and performance outcomes utilizing rigorous study designs is recommended.

Introduction

Biomedical informatics applications have been shown to improve efficiency and quality of health care delivery in management of diseases such as diabetes, cardiovascular disease, mental illness, and pediatric asthma, among others.¹⁻³ Clinical decision support (CDS) systems also have demonstrated improvements in practitioner performance.⁴⁻⁶ For instance, a systematic review by Garg et al. found that the majority of the reviewed electronic health record systems (EHRs) improved clinicians’ performance in drug prescribing and dosing, preventative care, and disease management.⁶ Likewise, Jaspers et al. concluded that CDS systems have a positive effect on provider performance regarding drug prescribing and preventive reminders.⁴

In the clinical setting, studies have found that poor nutritional support can contribute to impaired patient condition and therapeutic complications.⁷⁻¹⁰ For example, incidence of blood glucose (BG) dysregulation is associated with adverse outcomes and increased mortality.¹¹⁻¹² Alternatively, carefully monitored nutritional support providing adequate kilocalories and protein leads to lower rates of mortality and improved clinical outcomes in observational studies.¹³⁻¹⁸ Despite the benefits, nutritional monitoring can be a complex task, requiring management of multiple data inputs, mastering complex calculations, and consuming significant time from clinicians.¹⁹

Despite the growing body of research regarding health informatics technologies in medical patient management, no systematic reviews have been published describing the evidence for medical nutrition informatics applications. This paper aims to address that gap by summarizing the evidence of the effect of informatics applications containing CDS and computerized provider order entry (CPOE) features on 1) clinician workflow and error rate; 2) patient outcomes such as nutrient delivery, blood glucose management, malnutrition and weight loss; and 3) quality of nutrition support in various clinical settings. We use the term CDS to cover a broad variety of tools, from simple calculators to advanced algorithms. Where this use might cause confusion, we clarify the type of tool being discussed.

Methods

We conducted a systematic review of the peer-reviewed literature using three search databases: MEDLINE/PubMed, Scopus, and CINAHL (the Cumulative Index to Nursing & Allied Health Literature). Search criteria were designed broadly to include as many results as possible and employed keywords to search title, abstract, and MeSH (Medical Subject Heading) terms. A search was conducted on CINAHL using the keywords “nutrition AND informatics.” The following search terms were used for PubMed and Scopus databases: ("Nutrition Therapy"[Mesh] OR...
For all searches, we further limited relevant results to those published in the English language within the past ten years (January 1, 2004 and onward). Papers were accepted if they contained a clinical and a dietetic focus, an emphasis on CDS and/or CPOE applications, and human subjects of any age or disease state. Four reviewers independently assessed the set of returned titles and abstracts using those predefined inclusion criteria. Publications deemed irrelevant were excluded and duplicates were removed. Any discrepancies in the review process were resolved through unanimous consensus. We conducted a hand search of article bibliographies and related citations for additional sources. Upon completion of the initial abstracts search, we screened each paper in full for applicability, and identified final articles to be included in the systematic review. In total, sixteen studies were selected for analysis. The search process is illustrated in Figure 1.

**Figure 1.** Schematic of Search Strategies

**Results**

**General Overview**

A diverse set of locations and healthcare systems around the world are represented in this review. Of the sixteen included studies, three were conducted in Greece, two in the Netherlands, two in the United States, one in Switzerland, one in France, one in the Philippines, one in Norway, one in Denmark, one in Belgium, one in Austria, and one in Spain.

Of these studies, eleven examined the impact of a variety of CDS tools and CPOE systems on clinical patient outcomes and quality of nutrition care, two described features and validation of a computerized tool, four discussed effects on clinician workflow (for example, time saving features), and four assessed impact on rates of calculation and order error. Of the articles assessing the effect of informatics applications on patient care, five studies examined effects on BG values, related to use of computerized tools, four addressed energy or macronutrient delivery and achievement of prescribed nutritional support rate, one discussed unintentional clinical weight loss, and one studied recognition of malnutrition.
Systems varied regarding decision support and ordering features. Thirteen programs recommended nutrition and/or insulin interventions based on patient data, five contained computerized nutrition support ordering capabilities, five produced electronic alerts and reminders, three reported daily energy delivery totals from multiple sources, three conducted computerized patient screening, two produced feedback based on patient progress, and one developed automated meal plans based on patient needs.

The focus of these research articles generally fell within three major groupings: nutrition support for adults and neonates, BG management, and other nutrition applications. The design and duration of these studies varied widely among these three groups. Of the studies addressing impact of computerized interventions on patient outcomes, some featured a control group, though many did not. Others assessed features based on simulations, system descriptions, validation studies, or reductions in rates of calculation error.

**Adult Nutrition Support**

The ability to integrate and display nutritionally relevant patient data from multiple sources is a unique feature of computerized information systems. One CDS system synthesized data from delivered enteral or parenteral nutrition as well as macronutrient contributions from patient medications to provide a complete picture of nutrient delivery. Similarly, Berger et al. described one system able to synthesize energy totals from multiple sources, including glucose, lipids, and protein from medications and feeding, in addition to inputs from nutritional support. As a result of these data management features, clinicians responded well to the better graphical display of weight curves, energy balances, and total lipid, glucose, and protein deliveries with computerized systems. This display resulted in acceleration of nutrition support infusion rate when a deficit was observed. The assimilation of multiple data sources contributed to coordination and continuity of patient care when this information was available over prolonged time periods. Furthermore, some systems allowed the user to extract data to be analyzed by other computer programs for spreadsheet creation or statistical analysis.

Error rates were demonstrated to be lower in calculations performed by CDS in contrast to manual methods. Paschidi et al. developed a software tool for computation of parenteral nutrition prescriptions in adult patients based on validated estimated nutrient needs calculations. When compared to calculations completed without the assistance of a computer program, researchers observed a significant decrease in errors among a random selection of 21 patients. In general, use of computer-based nutritional support applications were shown to save time spent performing calculations.

Multiple studies employed use of computerized alerts, warnings, and feedback features to aid in selection and maintenance of nutrition support prescriptions. Decision support systems acted as a guide for the prescriber, monitoring delivered nutritional therapy, and warning of mismatched calculations and potential complications. Electronic warnings were implemented into nutrition support applications to identify values outside a normal range and/or in need of correction. Clinicians could take the initiative to seek out assistance, clicking on an “advice” button to receive recommendations for type and amount of nutritional formula. Alternatively, some feedback features were automatic. One system contained programmed informative messages intended to supplement clinician knowledge. Another sent the patient’s care team a daily email containing a comparison of delivered nutrition support to recommended calculations.

Patient outcomes were clearly and positively improved by informatics tools. The number of patients achieving early and adequate nutritional support increased significantly after the deployment of computer-aided support, in some cases nearly doubling, along with decreases in ICU (intensive care unit) length of stay after utilization of an “advice” button and daily feedback regarding achievement of nutrient delivery. In a study examining a total of 166 patients, higher numbers of critically ill patients received at least 80% of prescribed nutrition therapy for kilocalories (79% vs. 45%, p<0.001) and protein (37% vs. 3%, p<0.001) when a computer system was used for calculation of nutrition support prescriptions, with continued improvements over time. The same authors also reported that the group receiving computer input had increased percentages of patients prescribed at least 120% of recommended kilocalories. Berger et al. described the impact of an EHR on quality of nutritional support among 109 adult intensive care and burn patients. Though energy delivery was below target in both a unit containing a computerized system and one without, it was significantly higher and closer to desired level in the unit with computer support providing recommendations and feedback for discrepancies between calculated needs and actual energy delivery. Additionally, moderately and severely burned patients exhibited lower levels of weight loss when on the unit containing an EHR with these decision support features.
Improved energy delivery and avoidance of weight loss were not the only benefits observed in electronic CDS and CPOE systems. Llido reported that among 135,888 patients, malnourished individuals were recognized promptly based on a computerized screening program, leading to an increase in rates of referrals of these underweight patients to dietitians and nutrition support professionals from 37% to 100% . The percentage of critical care patients seen by nutrition support teams also increased from 10% to 99% . Computer systems used to calculate appropriate parenteral nutrition orders contributed to early identification of metabolic complications (25% increase in identification post-implementation). Additionally, one study reported a significant decrease in occurrence of ICU-acquired infection among a group of 95 patients whose nutrient needs were calculated using a computer program. However, the same study found no significant difference in mortality or length of ICU stay between the intervention and control group. Nutrition support findings are summarized in Table 1.

### Table 1. Adult Nutrition Support Applications

<table>
<thead>
<tr>
<th>Source</th>
<th>Population</th>
<th>System Features</th>
<th>Performance Outcomes</th>
<th>Patient Outcomes</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Llido, 2005</td>
<td>Hospital inpatients receiving NS</td>
<td>CDS, computerized patient screening</td>
<td>Data recording, nutrition support referrals, NSS patient care.</td>
<td>–</td>
<td>Entry of height/weight data in record increased from 30% to 90%. Referrals to NSS increased from 37% to 100%. Patients covered by NSS increased from 38.8% to 83%. Critical care patients seen by NSS increased from 10% to 99%.</td>
</tr>
<tr>
<td>Berger et al., 2006</td>
<td>Adult ICU patients receiving NS</td>
<td>CDS, integration of multiple data sources</td>
<td>Calculation time and completion</td>
<td>Energy delivery, weight loss</td>
<td>Computations were more complete and took less time in intervention group (2 min/patient vs. 11 min/patient, p=0.001). Proportion of postpyloric feeding days was higher in units with computer systems (p=0.003). Additionally, energy delivery was higher than control (84% goal vs. 41% goal, p=0.01) though both were under target. Among burn patients, computer managed group exhibited lower levels of weight loss (0.7 kg vs. 5.7 kg, p=0.034).</td>
</tr>
<tr>
<td>Paschidi et al., 2006</td>
<td>–</td>
<td>CDS, electronic alerts</td>
<td>Calculation time, rates of errors, identification of metabolic complications</td>
<td>–</td>
<td>Use of system resulted in 83% decrease in time required for calculations, 56% decrease in erroneous calculations, and 25% increase in early identification of metabolic complications.</td>
</tr>
<tr>
<td>Mirtallo et al., 2009</td>
<td>Inpatients receiving NS</td>
<td>CDS, CPOE, electronic alerts</td>
<td>–</td>
<td>–</td>
<td>Staff felt that the application improves continuity of care and saves time. Features included the ability to extract data to spreadsheets for analysis.</td>
</tr>
<tr>
<td>van Schijndel et al., 2009</td>
<td>Adult ICU patients receiving NS</td>
<td>CDS, integration of multiple data sources</td>
<td>–</td>
<td>Nutrient delivery</td>
<td>CDS system increased percentage of patients receiving adequate nutrition from 35.2% to 58.5% with a decrease in ICU length of stay.</td>
</tr>
<tr>
<td>Consell et al., 2013</td>
<td>Adult ICU patients receiving NS</td>
<td>CDS</td>
<td>–</td>
<td>Achievement of nutrition goals</td>
<td>No significant difference in duration of NS between groups. Computer group showed increased prescription of at least 120% of recommended kcal (12% vs 1% of patients, p=0.05). Calorie intake was higher with CIS (average 1793 kcal vs. 1508 kcal, p=0.01), as was nitrogen intake (11 g/day vs. 8 g/day, p=0.001). Computer group showed less frequent occurrence of ICU-acquired infection (59% vs. 41%, p=0.03), though there was no significant difference in mortality or length of ICU stay.</td>
</tr>
</tbody>
</table>

NS=nutrition support, CDS=computerized decision support, NSS=nutrition support services, ICU=intensive care unit, CPOE=computerized provider order entry

### Neonatal Nutrition Support

Premature and sick neonates frequently require nutrition support for adequate hospital recovery. Nonetheless, therapeutic errors can result in serious complications, so impacts on patient safety are routinely studied in neonatal nutrition support applications. A summary of the reviewed neonatal nutrition support applications is found in Table 2. Lehmann, Conner, and Cox researched an online CPOE system allowing for automated total parenteral nutrition (TPN) calculations in a neonatal ICU. Two years after implementation, an 89% reduction in total calculation errors was observed, amounting to a total of 1.2 errors per 100 orders vs. the initial rate of 10.8 errors per 100 orders. Likewise, Skouroliaikou et al. developed a computer program to assist in formulating and administering TPN to preterm and sick-term neonates. TPN orders prepared by the computer program contained no errors, compared to physicians’ orders, which had an error rate of 2.98%.
Further, CPOE systems in the neonatal ICU setting were met with enthusiasm from clinical staff. Results from staff-administered questionnaires rated computer systems as easier to learn and to use than paper systems, and as a better protection against errors.27 Changes in clinician workflow were also noted, with an average decrease from 7 minutes to 1 minute spent completing TPN orders, representing a reduction in clinicians’ workload.21

**Table 2. Neonatal Nutrition Support Applications**

<table>
<thead>
<tr>
<th>Source</th>
<th>Population</th>
<th>System Features</th>
<th>Performance Outcomes</th>
<th>Patient Outcomes</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lehmann et al.,24 2004</td>
<td>Neonatal ICU patients</td>
<td>CPOE, CDS, electronic alerts</td>
<td>Rates of error, calculation time</td>
<td>–</td>
<td>There was a 100% reduction in incomplete TPN order forms at 2 years. An 89% reduction in errors (including adherence to osmolality guidelines, calculations, and other knowledge deficiencies) was observed (1.2 errors per 100 orders vs. 10.8 errors per 100 orders at baseline).</td>
</tr>
<tr>
<td>Skouroliakou et al.,23 2005</td>
<td>Pre-term and sick-term neonates</td>
<td>CPOE, CDS, electronic alerts</td>
<td>Rates of error, calculation time</td>
<td>–</td>
<td>Manual calculations contained an error rate of 2.98% vs. 0% error rate with computer calculations (p&lt;0.001). The new system resulted in an average of 1 minute per calculation vs. 7 minutes without the computer (p=0.05).</td>
</tr>
</tbody>
</table>

ICU=intensive care unit, CPOE=computerized provider order entry, CDS=computerized decision support

**Blood Glucose Management**

A substantial number of studies investigated CDS and CPOE systems on BG management during nutrition therapy, as shown in Table 3. In general, informatics tools were utilized to balance delivery of nutrition with management of BG levels. Hoekstra demonstrated a viable method for initiating nutrition support while managing target BG values in 23 predominantly surgical ICU patients.24 By utilizing an incremental step-up method of PN initiation in addition to a CDS system for continuous monitoring of insulin infusions, adequate caloric and nutrient allocation was achieved within 24 hours while consistently maintaining BG levels under acceptable parameters for the greatest majority of measurements in patients.25 Similarly, another CDS system in a neurotrauma ICU allowed for provision of 93.5% of estimated resting energy expenditure for adult patients (n=6) vs. 129.5% in the control group (n=6) while also maintaining BG levels at an average lower than that of the control group.22

In addition to adequate nutrient delivery, CDS aided in the management of BG levels. Pachler et al. described a computer algorithm for management of mechanically ventilated adult ICU patients.36 Taking into account current BG concentration, insulin dosage, and carbohydrate content of nutrition support formula, calculations were determined for insulin infusion rate and timing of next BG sample. Under these conditions, the intervention group exhibited lower average BG concentrations and hyperglycemic index levels compared to the control group.36 Among these same patients, the total carbohydrate administration rate was no different between intervention and control group. Meyfroid et al. found that a glucose alert system for adult ICU patients employing pop-up alerts in response to abnormal values decreased mean BG values, decreased hyperglycemic index levels, decreased glycemic penalty index levels, and decreased proportion of measurements within the hyperglycemic range.35 There were no significant differences in hypoglycemic index, or proportion of measurements in the hypoglycemic range, however.35 Another study found that a CDS and CPOE system for neonatal patients had no effect on instances of hypo- or hyperglycemias pre- and post-intervention, despite a modest reduction in time for simple calculations, and major reduction in time for complex calculations.27

**Other Nutrition Applications**

Apart from parenteral and enteral nutritional support informatics applications, several programs have been developed to assist with patient screening and oral feeding, as illustrated in Table 4. de Ulibarri et al. implemented a nutritional risk screening and referral system based on laboratory parameters in an adult hospital population not undergoing aggressive therapeutic procedures or suffering from severe diseases.35 Compared to validated measures, the system had a screening sensitivity of 92.3% and a specificity of 85%.35 In a related study, Fossum et al. found that a CDS system used for risk screening in nursing homes for the elderly resulted in significantly reduced rates of malnourishment among a baseline of 491 individuals when compared to institutions without such a program.31 However, the same study found no effect on rates of pressure ulcer formation,31 even though nutritional status is considered an important predictor of the condition.
### Table 3. Blood Glucose Management Applications

<table>
<thead>
<tr>
<th>Source</th>
<th>Population</th>
<th>System Features</th>
<th>Performance Outcomes</th>
<th>Patient Outcomes</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pachler et al., 2008</td>
<td>Adult ICU patients</td>
<td>CDS</td>
<td>–</td>
<td>Mean BG, hyperglycemic index</td>
<td>Median BG decreased in computer group (5.9 mMol/L [106.3 mg/dL] vs. 7.4 mMol/L [133.3 mg/dL] control, p&lt;0.001). Hyperglycemic index decreased in computer group (0.4 mMol/L [7.2 mg/dL] vs. 1.6 mMol/L [28.8 mg/dL] control, p&lt;0.001). Total carbohydrate administration rate was not significantly different between groups.</td>
</tr>
<tr>
<td>Hoekstra et al., 2010</td>
<td>Adult ICU patients</td>
<td>CDS</td>
<td>–</td>
<td>Caloric intake, BG levels</td>
<td>With CDS system, caloric intake goal was reached at median 18 hours, maximum 24 hours for all patients. The greatest majority (95.5%) of BG measures were below 10 mMol/L (180.2 mg/dL).</td>
</tr>
<tr>
<td>Meyfroidt et al., 2011</td>
<td>Adult ICU patients</td>
<td>CDS, electronic alerts</td>
<td>–</td>
<td>Mean BG, hyperglycemic index, glycemic penalty index, number of hypoglycemic events</td>
<td>System resulted in a decrease in mean BG level from 112 mg/dl to 110 mg/dL (p=0.002), a decrease in glycemic penalty index from 20 to 19 (p=0.029), and a decrease in hyperglycemic index from 10 mg/dl to 9 mg/dL (p&lt;0.04). There was a decrease in hypoglycemic episodes from 6.5% to 4.0% (p&lt;0.04).</td>
</tr>
<tr>
<td>Pielmeier et al., 2012</td>
<td>Adult ICU patients</td>
<td>CDS</td>
<td>–</td>
<td>Mean BG, caloric intake</td>
<td>CDS system increased percentage of BG measurements within 5-8 mMol/L (50.1-144.1 mg/dL) [76% intervention vs. 51% control, p=0.05]. Mean BG in CDS group was 7.0 mMol/L (126.1 mg/dL) vs. 8.0 mMol/L (144.1 mg/dL) control, p=0.05. Mean caloric intake in computer group was 93.5% of estimated needs vs. 129% in control. Insulin infusion rates did not significantly differ between groups.</td>
</tr>
<tr>
<td>Maat et al., 2013</td>
<td>Neonatal ICU patients</td>
<td>CPOE, CDS, multiple data source integration</td>
<td>Prescribing time</td>
<td>Incidence of hypo- or hyperglycemic events</td>
<td>The computer system resulted in a time reduction of 16% for simple calculations (basic glucose dosing guidance) and reduction of 60% for complex calculations (total glucose delivery from all sources including nutrition support and drug vehicles). There was no significant difference in pre- and post-CPOE mean incidence of hypo- or hyperglycemicas.</td>
</tr>
</tbody>
</table>

ICU=intensive care unit, CDS=computerized decision support, BG=blood glucose, CPOE=computerized provider order entry

### Table 4. Other Nutrition Applications

<table>
<thead>
<tr>
<th>Source</th>
<th>Population</th>
<th>System Features</th>
<th>Performance Outcomes</th>
<th>Patient Outcomes</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fossum et al., 2011</td>
<td>Elderly nursing home residents</td>
<td>Computerized screening, CDS</td>
<td>–</td>
<td>Risk for a prevalence of pressure ulcers and malnutrition</td>
<td>The CDS system significantly reduced rates of malnourishment (defined as inadequate nutritional status) when compared to institutions without such a program (9% decrease in malnourished residents in intervention vs. 3.9% increase control, p&lt;0.05). There was no significant effect on pressure ulcer formation.</td>
</tr>
<tr>
<td>Skourolaikou et al., 2009</td>
<td>Adult hospital patients not receiving NS</td>
<td>Automated meal planning</td>
<td>Rates of error, data recording, calculations</td>
<td>–</td>
<td>The system resulted in a reduction in errors occurred during data recording, calculation of daily requirements, and menu planning (88% decrease, p&lt;0.05). Dietitians experienced a minimization of time spent on menu planning after program implementation (68% decrease, p&lt;0.05). There was no significant effect on days of hospitalization.</td>
</tr>
<tr>
<td>Ignacio de Ulibarri et al., 2005</td>
<td>Adult hospital patients</td>
<td>Computerized screening, CDS</td>
<td>Validation with accepted screening methods</td>
<td>–</td>
<td>The CDS had a screening sensitivity of 92.3% and a specificity of 85% when compared to validated measures.</td>
</tr>
</tbody>
</table>

CDS=computerized decision support, NS=nutrition support
Skouroliakou et al. published an implementation brief describing an application of nutrition informatics in oral nutrient delivery for hospitalized patients. This system was developed to prepare dietary prescriptions, calculate nutritional requirements, store patients’ dietary records, produce automatic daily menus, and generate reports on food lists and menu costs for patients receiving nutritional support by mouth. As a result of that implementation, a reduction in errors occurred during data recording, during calculation of daily requirements, and during menu planning. Additionally, authors reported a minimization of time spent by dietitians on menu planning after program implementation. For patients, no significant effect on days of hospitalization was found related to use of this tool.

Discussion

This systematic review examined a wide range of medical CDS and CPOE informatics applications in clinical dietetics. Overall, the benefits described in the reviewed articles were numerous. Specifically, improved patient outcomes included lower rates of weight loss, increased nutrient delivery, fewer acquired infections, more stable BG values, and lower rates of malnutrition, in addition to increased referrals to nutrition support teams. Furthermore, improved clinician workflow and fewer calculation errors were noted in multiple studies. Researchers reported high rates of acceptance of CDS and CPOE systems among clinical staff. However, studies produced conflicting results regarding the influence of computerized interventions on length of hospital stay, with a tendency toward no effect. Only one study assessed mortality outcomes; here, researchers found no significant effect. Taken together, these results suggest that use of informatics applications contribute to an improvement in metabolic management, but they offer a conflicting and incomplete picture of overall patient outcomes.

Although reviewed studies showed significant improvements in error reduction among adult and neonatal nutrition support ordering, in only one study was the absolute error number reduced to zero. One might argue that any level of error is unacceptable, particularly among system-generated calculations. It is important to note the synergistic relationship between CDS and CPOE systems and clinical care providers. Though CDS systems may generate automated calculations, this result is dependent on completeness of input patient data as well as clinicians’ adherence to suggested electronic alerts and reminders. Regardless of the demonstrated value of CDS systems, there is no substitution for qualified human clinical judgment in delivery of healthcare. Even the most well designed system might not produce perfect results without accurate human knowledge and attention to detail. Clinician oversight in regards to prescription and delivery of adult and neonatal nutrition support is always necessary to ensure best patient outcomes.

Interestingly, there was a more effective reduction of errors in the neonatal nutrition support systems than in adult nutrition support systems. It could be that neonates and sick term infants are more likely to have a similar set of clinical needs and treatments than is seen in the large variety of adult patients of various disease states. However, the authors are more inclined to think that user interface may have the greatest impact on error reduction. For example, the system described by Skouroliakou generates a set of yes/no questions for the user to answer to determine parameters for preparation of TPN. This type of simplification of input may clarify the most important clinical parameters when prescribing nutrition support. Particularly in hectic environments such as an adult or neonatal ICU, simplified but thorough CDS systems will be the most useful in assisting nutrition support prescription.

Some tools may be better suited to particular clinical situations than others. In particular, the reviewed studies have demonstrated that clinical applications for TPN ordering are well suited for delivery of neonatal care. Additionally, blood glucose management systems have a potential application in all clinical areas, from ICU to medical floors to neonatal care. Likewise, nutrition informatics applications have a place in multiple types of hospitals; for rural hospitals, CDS systems may supplement and inform clinical decision making when specialized nutrition professionals may not be readily available. Although every clinician can benefit from use of CDS systems, CPOE is best suited for ordering providers such as physicians, physician’s assistants, and pharmacists. EHRs designed with multiple CDS features will fulfill the greatest number of needs, with access privileges for clinicians specific to need.

Although this review discusses nutrition informatics applications in clinical practice, the potential of informatics goes beyond inpatient use. Recent technological innovations targeting patient self-management, especially through mHealth apps, suggest clinical efficacy. Oenema et al. reported that an Internet-delivered computer-tailored lifestyle intervention program resulted in lower saturated fat intake based on a validated food-frequency questionnaire (p<0.01). In addition to modifications to patient behavior, computerized systems have the potential to impact outcomes such as weight management. For example, research has suggested that online “feedback” features such as
progress charts and physiological calculators are the best predictor of weight loss in internet-based weight control programs.\textsuperscript{38}

There are also implications for chronic disease management. Albisser described a Web-based graphical user interface designed to provide decision support regarding medication dosing and lifestyle factors such as diet and exercise for individuals managing insulin-dependent diabetes.\textsuperscript{25} Any actions taken by the user were documented by the system and linked back to an online registry, where these data were available for review and/or intervention by the patient’s provider. This application suggests the potential integration of patient-facing technologies and clinical care coordination,\textsuperscript{39} particularly in diseases requiring joint patient-provider supervision. These findings seem to suggest that decision support features in online programs and applications designed for patient self-management might contribute to positive clinical outcomes in a variety of diseases.

This study had several limitations. First, results included only those published in English, potentially excluding meaningful studies written in other languages. Second, this paper is necessarily limited by the methods used in the primary research reviewed. The largest majority of the studies reviewed in this paper employed a quasi-experimental study design (n=10, 62.5\%). Several other papers were published as a system description,\textsuperscript{20,26} proof of concept study,\textsuperscript{24} or validation study.\textsuperscript{25} Only two studies employed a controlled trial method, one with group randomization,\textsuperscript{32} and one an “open” design.\textsuperscript{34} As such, overall results should be considered preliminary findings in the area of nutrition informatics applications and interpreted with some caution. Future research should consider more rigorous methods, including randomized controlled trials utilizing large sample sizes. Finally, a meta-analysis was impossible due to significant differences in reported outcomes. However, the positive results described herein suggest that further research examining the benefit of informatics interventions, specifically those incorporating CDS and/or CPOE, will strengthen the case for nutrition informatics applications in the clinical setting.

Furthermore, it is important to note the stage of development of each informatics approach when assessing reported results. Though studies showed positive outcomes, not all computer systems had been implemented in an actual clinical patient-care setting.\textsuperscript{20} It is possible that systems that look promising in the early stages may not translate to practical use in patient care. Therefore, studies set in a clinical venue are needed for further conclusions. Although the studies included in this review provide examples of informatics tools used in nutritional management, very few specifically addressed patient health outcomes. The effect of CDS and CPOE systems on length of hospital stay, weight loss, wound healing, and mortality represent important markers of nutrition care efficacy. Overall, the impact of informatics on patient outcomes remains an understudied area of research in all medical informatics applications, including those used for nutrition care.

Use of informatics applications is relevant for Registered Dietitians as well as other nutrition care professionals. Specifically, the integration of informatics systems into clinical practice may increase practitioner efficiency as well as improve the quality of nutrition care. For example, Skouroliakou described an informatics application developed to specifically improve efficiency and performance of dietitians in meal planning and menu assessment.\textsuperscript{22} Lildo demonstrated that use of an informatics system resulted in increased involvement of clinical dietitians in all aspects of patient care, including referrals to nutrition support services.\textsuperscript{30} Additionally, in the application described by Mirtallo et al., Registered Dietitians were consulted throughout the development and initiation process for system evaluation and feedback, ensuring optimized nutrition care functionality.\textsuperscript{26}

For Registered Dietitians, CDS systems present an opportunity to enhance quality of patient care as well as number of malnourished and critically ill patients seen by nutrition care practitioners. Many healthcare institutions implement a screening procedure for patients who qualify for a dietitian referral; often dietary technicians or the dietitians themselves complete this screening task. Automated systems such as the one described by Lildo\textsuperscript{22} may be helpful in identifying these at-risk patients and initiating nutrition care contact sooner while reducing the workload of hospital personnel. Among critical care patients, nutritional status is often an area of concern as many patients are unable to meet nutrition needs for at least some portion of their hospital stay depending on the severity of their medical status. Increased referrals to nutrition services likely represent an early identification of potential need for alternative support in at-risk patients especially susceptible to malnutrition. When tailored to reflect the screening parameters set forth for a particular hospital, systems developed to correctly identify patients in need of nutrition intervention ensure that the standard of nutrition care is upheld.

We would argue that nutrition professionals can use the research from this review to advocate for informatics systems to assist in delivery of nutrition support and metabolic management, to improve calculation accuracy when writing prescriptions, and to assist in screening and early identification of nutritional concerns. Furthermore, these
applications make possible the direct integration of evidence-based recommendations for clinical nutrition care into ordering systems, including standardized processes for nutrition support ordering, delivery, and monitoring. In conclusion, acceptance of nutrition informatics applications is increasing in inpatient and outpatient healthcare, both nationwide and worldwide. Appropriate nutrition monitoring and support is an integral part of ensuring hospital recovery and positive patient outcomes. Given this, computerization of the nutrition care process in particular could impart significant benefits for patient outcomes, practitioner performance, and institutional cost. Further research is recommended to provide evidence for integrating informatics applications into routine clinical nutrition practice.

References


Christopher Ochs, MS1, Ling Zheng, MS1, Huanying Gu, PhD2, Yehoshua Perl, PhD1, James Geller, PhD1, Joan Kapusnik-Uner, PharmD3, Aleksandr Zakharchenko, BS1
1New Jersey Institute of Technology, Newark, NJ; 2New York Institute of Technology, New York, NY; 3First Databank, Inc., San Francisco, CA

Abstract

The National Drug File – Reference Terminology (NDF-RT) is a large and complex drug terminology. NDF-RT provides important information about clinical drugs, e.g., their chemical ingredients, mechanisms of action, dosage form and physiological effects. Within NDF-RT such information is represented using tens of thousands of roles. It is difficult to comprehend large, complex terminologies like NDF-RT. In previous studies, we introduced abstraction networks to summarize the content and structure of terminologies. In this paper, we introduce the Ingredient Abstraction Network to summarize NDF-RT's Chemical Ingredients and their associated drugs. Additionally, we introduce the Aggregate Ingredient Abstraction Network, for controlling the granularity of summarization provided by the Ingredient Abstraction Network. The Ingredient Abstraction Network is used to support the discovery of new candidate drug-drug interactions (DDIs) not appearing in First Databank, Inc.'s DDI knowledgebase.

Introduction

We present a new method for the discovery of candidate drug-drug interactions (DDIs). Conceptually, this method is based on comparing a large commercial knowledgebase of DDIs with small groups of drugs where all members of one group contain similar ingredients. These groups of drugs were derived from an independent drug terminology, the National Drug File – Reference Terminology (NDF-RT) [1]. The challenge is that the NDF-RT in its source format does not provide appropriate groups of drugs. Finding such groups by hand is difficult, because the NDF-RT is composed of approximately 43,000 concepts connected by 67,000 IS-A roles and 73,000 other roles, e.g., has ingredient and has mechanism of action. Thus, we are using algorithms, based on a well-developed body of work on abstraction networks (AbNs) [2], to derive small groups of similar drugs. Intuitively, if most but not all members of such a group have known drug-drug interactions according to a DDI knowledgebase, then the remaining group members should be reviewed by an expert to determine why they don’t have (known) DDIs. The challenge is that none of our previously developed AbNs could be applied to the NDF-RT. In this paper, we present a new AbN, expressly developed for the NDF-RT and for the purpose of discovering candidate DDIs.

NDF-RT uses a description logic-based concept model to define drugs according to various aspects, e.g., active chemical ingredient, mechanism of action, physiologic effect, therapeutic intent, and dosage form. Each aspect is represented by a separate concept hierarchy and the aspects of each drug are expressed via roles. For example, the drug concept Aspirin/Caffeine has two has ingredient roles, to Aspirin and to Caffeine.

Abstraction networks (AbNs) [2] are compact, visual terminology summaries, based on grouping “similar” terminology elements together into groups. (The meaning of “similarity” is terminology-dependent.) One type of AbN that was previously applied to description logic-based terminologies is the partial-area taxonomy [3], thus one would expect that this AbN could be applied to the NDF-RT. However, due to NDF-RT’s highly specialized structure, our previously developed methods for deriving AbNs are not applicable to it, because the majority of NDF-RT’s hierarchies have no roles emanating from their concepts. The existence of such roles is the basis for our previously developed AbNs for description logic-based terminologies.

In this paper, we describe the derivation of a new kind of AbN called an Ingredient Abstraction Network (IAbN), which summarizes NDF-RT’s chemical ingredients and their associated drug concepts. The NDF-RT distinguishes between different dosage forms of the same drug. However, for the purpose of the discovery of potential DDIs we can safely ignore such distinctions in our algorithms. The resulting IAbN turned out to be not sufficiently compact, which is a desirable feature. Thus, we developed a secondary abstraction mechanism (“Aggregate IAbN”) that creates an IAbN with fewer groups and each group stands for a collection of groups from a “complete” IAbN (i.e., a non-aggregate IAbN). The Aggregate IAbN derivation algorithm is parameterized to allow users control over the degree of summarization that is achieved.
Background

NDF-RT is a formal representation of the VHA National Drug File (NDF) [4], which is a drug classification hierarchy used to group orderable drug products into one of 579 drug classes. NDF is used to support VHA’s clinical applications. NDF-RT’s Mechanism of Action, Physiologic Effect and Chemical Ingredients hierarchies were created by matching VHA drug ingredient names to terms from the National Library of Medicine (NLM)’s Medical Subject Headings (MeSH) [5]. Specifically, the Chemical Ingredients (CI) hierarchy was derived from MeSH’s Chemicals and Drugs Category. The Mechanism of Action and Physiologic Effect hierarchies were also initially created based on MeSH [6] and then extended. NDF-RT organizes concepts around the Pharmaceutical Preparations (PP) hierarchy, the largest one in NDF-RT, with 25,093 concepts (July 2014 version). Besides IS-A roles, PP concepts can have roles to concepts in other hierarchies to define drugs according to their various aspects. For example, the drug concept Aspirin in PP has the role has ingredient to Aspirin in CI, the second largest hierarchy in NDF-RT with 10,118 concepts.

An abstraction network [2] (AbN) is defined as a compact network of nodes that summarize groups of “similar” concepts in a terminology to capture its “big picture.” The definition of “similar” is based on the type of terminology and the specific kind of abstraction network. AbN nodes are organized in a hierarchy derived from the hierarchical IS-A links of the terminology. In previous work, we have derived various [3, 7-10] kinds of abstraction networks for many different terminology systems, e.g., the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) [11], the National Cancer Institute thesaurus (NCIt) [12], and the Gene Ontology (GO) [13]. These AbNs were shown successful in supporting the identification of concepts with a high likelihood of errors. For a review of different kinds of AbNs and their properties, see Halper et al. [2].

Extensive research has been done on NDF-RT, e.g., on its content coverage, the adequacy of representation, drug normalization and classification. Rosenbloom et al. [14] investigated the adequacy of the representation of the Physiologic Effect hierarchy. The results suggested that the concepts in the Physiologic Effect hierarchy are appropriate for medications. Carter et al. [15] studied drug class names from three sources to understand how drugs were classified, and evaluated NDF-RT’s semantic coverage. They found that NDF-RT can cover more than 90% NDF drug categories. Pathak et al. [16] evaluated the applicability of RxNorm [17] and NDF-RT for classification of medication data extracted from electronic health records (EHRs). Their study demonstrated that the two terminologies can be used together for drug classification.

Methods

Definition: An Ingredient Abstraction Network (IAbN) is an AbN where the nodes summarize (1) the ingredients in the Chemical Ingredient hierarchy and (2) those drug concepts in the PP hierarchy that have no dosage information but that do have at least one has ingredient role to a drug ingredient in the Chemical Ingredient hierarchy.

Refer to Figure 1 for the following definitions that will be used throughout the whole paper. We distinguish between four types of concepts in the Chemical Ingredients (CI) hierarchy. (1) A drug ingredient in CI is the target of has ingredient role(s) from concepts in the Pharmaceutical Preparations (PP) hierarchy. Drug ingredients are chemical ingredients that are used in prescription drugs. (2) A classification ingredient in CI is a concept that “organizes” other drug ingredients. It has drug ingredients as children. (3) A dual use ingredient in CI is both a drug ingredient and a classification ingredient. Note that this is “dual use” in the terminology, not for prescription. A classification ingredient that is not also a drug ingredient is called a strict classification ingredient. (4) Some concepts in CI are neither a drug ingredient nor a classification ingredient. No specific name was chosen for such concepts, as they are not used in this work. The right side of Figure 1(a) illustrates these definitions for an excerpt of 12 CI concepts.

The design of an AbN for the CI hierarchy poses a challenge for several reasons: (1) A lack of roles emanating from CI concepts prevents the derivation of AbNs called partial-area taxonomies [3]. (2) The need to distinguish between drug ingredients and classification ingredients is further complicated by the dual use of many CI concepts. (3) There is a need to summarize the drug concepts, which in NDF-RT are parts of the PP hierarchy, according to their ingredient concepts in CI. (4) To obtain a “big picture” of the classification ingredients, one needs to identify distinct groups of drugs and organize them in a way that supports DDI discovery. (5) A method is needed that allows a user to control the granularity of summarization, so that s/he is not overwhelmed.

IAbN derivation begins with identifying all of the drug concepts in the PP hierarchy with a has ingredient role but with no has_DoseForm role. PP concepts with dosage information are ignored since an ancestor concept, typically a parent (a PP generic drug ingredient) introduces the has ingredient role, which is inherited to such concepts. All of the PP concepts in Figure 1(a), except Pharmaceutical Preparations, have one has ingredient role to a concept in
the CI hierarchy. Different drug concepts in the PP hierarchy can have a has ingredient role to the same CI concept, e.g., both Aspirin and Acetylsalicylic acid have the ingredient Aspirin. PP concepts may also have multiple has ingredient roles, e.g., Aspirin/Caffeine has has ingredient roles to both Aspirin and Caffeine.

**Figure 1.** (a) An excerpt of concepts from NDF-RT’s Pharmaceutical Preparations (PP) and Chemical Ingredients (CI) hierarchies. On the left, drug concepts in the PP hierarchy with no dosage information have a shaded background. On the right, seven drug ingredients are outlined in pink and five classification ingredients have a pink background. Two concepts, Aminosalicylic acid and Warfarin, are both drug ingredients and classification ingredients, i.e., they are dual use ingredients. Ethyl Bicoumacetate is neither a drug ingredient nor a classification ingredient. (b) CI revisited: Drug ingredients (not shaded) and their has ingredient roles (shaded). Each drug ingredient is color-framed according to its lowest common ancestor classification ingredient according to its lowest common ancestor classification ingredient. (c) The final IABN for the excerpt of Figure 1(a). Ingredient groups are shown as white boxes that are labeled with the name of the lowest common ancestor from 1(b). Also shown are the total number of ingredient concepts summarized by the group, and the total number of drug concepts [with no dosage information in the PP hierarchy] with has ingredient roles pointing to the CI hierarchy. Child-of links between ingredient groups are shown as upward directed arrows.

Next, drug ingredients (see definition above) are identified by collecting the target concepts of all the has ingredient roles. Classification ingredients (again, see definition above) are identified by analyzing the parent concept(s) of each drug ingredient. Additionally, for each drug ingredient, the lowest ancestor(s) that are a strict classification ingredient(s) are identified, with the intention of finding lowest common ancestor classification ingredients for groups of drug ingredients. For example, for the Aspirin CI concept, the lowest ancestor that is a strict classification ingredient is Salicylates. Salicylates is the lowest common ancestor for Aspirin and Magnesium salicylate. For Warfarin Sodium its parent concept, Warfarin, is a classification ingredient but it is also a drug ingredient (i.e., it is dual use). Thus, the lowest ancestor of Warfarin Sodium that is a strict classification ingredient is Warfarin’s parent, 4-Hydrocoumarins. Many CI hierarchy concepts have multiple parents, thus, a given drug ingredient may have more than one lowest ancestor which is a strict classification ingredient.

Drug ingredients are grouped together according to their lowest common ancestor(s) that are strict classification ingredients. For example, Aspirin and Magnesium salicylate both share Salicylates as lowest common ancestor. Similarly, Warfarin and Warfarin Sodium share 4-Hydrocoumarins as a lowest common ancestor. Figure 1(b) models the right side of 1(a) and shows the drug ingredient groups induced by the lowest common ancestors.
Next, each strict classification ingredient is made into a root for its ingredient group. Thus Salicylates becomes the root of the group with Aspirin and Magnesium salicylate in it. The CI root concept, Chemical Ingredients, is also a root. Figure 1(c) shows how roots stand in for their groups. The line “2 Ingredients” under Salicylates indicates how much information is summarized. Ingredient groups are not disjoint; drug ingredients with multiple parents may be summarized by multiple ingredient groups. With this we have achieved a summary of the “right side” (the Chemical Ingredients) of Figure 1(a). In the next step, we include information from the left (PP) side into Figure 1(c).

For each ingredient group, the PP drug concepts that have a has ingredient role to a drug ingredient in the ingredient group are identified. For example, the Aspirin and Acetylsalicylate Sodium drug concepts in PP both have Aspirin in CI as the target of their has ingredient roles. The Aspirin drug ingredient belongs to the Salicylates ingredient group, thus, the Aspirin and Acetylsalicylate drug concepts from PP are also summarized by the Salicylates ingredient group. This is expressed by the line “3 Drugs” under Salicylates in 1(c). (The third is Magnesium salicylate). Since ingredients may belong to multiple ingredient groups, a given PP drug concept may be in multiple ingredient groups.

Within the IAbN, ingredient groups are organized into a hierarchy according to child-of links derived from the underlying IS-A hierarchy. An ingredient group A is a child-of another ingredient group B if A’s root has B’s root as an ancestor and there are no other roots on any path from A’s root to B’s root. An ingredient group may be child-of multiple ingredient groups. In the visualization of an IAbN it is necessary to organize the ingredient groups in a way that helps the summary reflect the “big picture.” Thus, ingredient groups are organized into color coded levels according to the length of the longest child-of path to the root ingredient group (Chemical Ingredients). Figure 1(c) shows the IAbN derived from the NDF-RT excerpt in Figure 1(a).

We note that there are CI concepts that are neither drug ingredients nor classification ingredients, e.g., Ethyl Bicoumacetate. (See case (4) in the above definitions and Figure 1(a).) This occurs when the drug ingredient is modeled in CI but no PP drug concept has a has ingredient role to this drug ingredient. For the current research, such concepts are not summarized by any ingredient group and are not considered part of the IAbN. In the Discussion we propose methods for extending the IAbN to include these concepts. Additionally, we note that drug concepts with dosage information are not associated with an ingredient group. However, the ingredient group(s) of these drug concepts can be identified via their parents, which are grouped into at least one ingredient group.

**Aggregate Ingredient Abstraction Network (Aggregate IAbN)**

One significant issue we encountered when deriving the IAbN for the complete CI hierarchy (“complete IAbN”) was its large size. While the complete IAbN is significantly smaller and less complex than the underlying CI hierarchy, there are still too many nodes, many of which “summarize” only one drug ingredient. To improve the efficacy of the IAbN to function as a summary, we now introduce a parametric method for controlling the granularity of summarization. This secondary summarization approach is based on the following heuristic: An ingredient group that summarizes a relatively large number of drug ingredients is more important within the CI hierarchy than an ingredient group that summarizes relatively few drug ingredients.

However, controlling the granularity of summarization provided by the IAbN requires more than just hiding small ingredient groups. Simply hiding these groups leads to a loss of information and inconsistencies in the child-of hierarchy. To obtain a more compact secondary summary of the CI hierarchy’s content we have developed a parametric IAbN called an Aggregate Ingredient Abstraction Network (“Aggregate IAbN”) that aggregates “small” ingredient groups into their larger direct ancestor ingredient groups.

Given a bound b (a natural number), an Aggregate IAbN is derived in the following way. Starting from the root ingredient group, Chemical Ingredients, the hierarchy of ingredient groups is traversed downwards using a Topological Sort; that means, an ingredient group is processed only after all of its parent ingredient groups have been processed. We define a) aggregate ingredient groups, b) removed ingredient groups, and c) regular ingredient groups. A removed ingredient group is an ingredient group in the complete IAbN with fewer than b ingredients. The root ingredient group, Chemical Ingredients, is by definition not a removed ingredient group. Approximtely speaking, an aggregate ingredient group is an ingredient group that summarizes itself and one or more removed ingredient groups. A regular ingredient group is not changed when going from a complete IAbN to an Aggregate IAbN. Thus the Aggregate IAbN consists of aggregate ingredient groups and regular ingredient groups.

More precisely, in an Aggregate IAbN, a removed ingredient group i is included into an aggregate ingredient group a if i is a descendant of a in the complete IAbN and there is no other aggregate ingredient group on any child-of path from i to a. Removed ingredient groups are “hidden” in the Aggregate IAbN. Consider the IAbN excerpt in Figure 2 and a bound b=10. Chemical Ingredients has two child ingredient groups with fewer than ten drug ingredients.
(Glycosides and Heterocyclic Compounds) and several further descendants with fewer than ten drug ingredients (e.g., Glucosides and Piperidones). Some of these further descendant ingredient groups may be direct descendants of other ingredient groups that have more than ten drug ingredients, e.g., Piperidones is child-of Piperidines, which summarizes 47 ingredients.

Consider the two removed ingredient groups that are children of Chemical Ingredients in the IAbN of Figure 2(a), with \( b = 10 \): Glycosides and Heterocyclic Compounds. There are no intermediate aggregate ingredient groups, so these children are included into Chemical Ingredients in the Aggregate IAbN of Figure 2(b). Furthermore, the removed ingredient group Glucosides is a child of Glycosides. The closest aggregate ingredient group above Glucosides that is not a removed ingredient group, via the path through Glycosides, is Chemical Ingredients. Thus, Glucosides is included into the Chemical Ingredients aggregate ingredient group as well. Note: \( I = 14 \) in Chemical Ingredients in Figure 1(b), summarizing the removed ingredient groups, while \( I = 3 \) in Chemical Ingredients in 1(a).

**Figure 2.** (a) An excerpt of 13 ingredient groups from the complete IAbN. Child-of links are shown as arrows. Removed ingredient groups (fewer than ten ingredients) have a red background. (b) The Aggregate IAbN (\( b = 10 \)) for (a). Aggregate ingredient groups are shown as rounded-corner white rectangles, labeled with the numbers of ingredients, drugs and groups they summarize. (\( I = \) Ingredients; \( D = \) Drugs; \( G = \) Groups that were removed). Regular ingredient groups are white rectangles. They have the same \( I \) and \( D \) value in (a) and in (b).

Aggregate ingredient groups are not necessarily disjoint in terms of which removed ingredient groups they summarize. There may be two or more paths to different aggregate ingredient groups that satisfy the above condition. In Figure 2(a) Pipecolic Acids is a child-of Piperidines and Acids, Heterocyclic. On the path from Pipecolic Acids to Chemical Ingredients via Acids, Heterocyclic there are only removed ingredient groups. Thus, in the Aggregate IAbN of Figure 2(b), Pipecolic Acids is summarized by both Piperidines and Chemical Ingredients.

The child-of links between ingredient groups in the Aggregate IAbN are established according to the ingredient group hierarchy of the complete IAbN. If an aggregate ingredient group or a regular ingredient group is a child-of a removed ingredient group, then its child-of link(s) in the Aggregate IAbN go to the aggregate ingredient groups above the removed ingredient group. In other words, the child-of link is redirected to the aggregate ingredient group that summarizes the removed ingredient group.

**IAbN Supporting the Discovery of New Candidates for Pharmacodynamic-based DDIs**

One application of the IAbN is the discovery of new candidate drug-drug interactions (DDIs) not appearing in a DDI knowledgebase. Given is a set of DDI rules in the form (Drug1, Drug2, Clinical Consequence), such that Drug1 and Drug2 can be coded by NDF-RT drug ingredient concepts. In pharmacology, drugs with the same chemical ingredients tend to have similar DDIs [18]. By reviewing the DDIs associated with the drug ingredients in an IAbN ingredient group one may uncover new candidate pharmacodynamic-based DDIs.

To illustrate this process, we use an example from First Databank, Inc.’s DDI knowledgebase [19]. Consider the 18 drug ingredients in the Aggregate IAbN’s Salicylates ingredient group (Figure 5(b)), 13 of which appear in FDB’s DDI knowledgebase. The DDI interactions between ten of these salicylates and seven anticoagulant drugs are “Avoid concurrent use when possible” (AVD) and “Increases the effect of latter drug” (INL), for a total of 70 DDIs between these two groups. However, three extra Salicylates (balsalazide, mesalamine, and salsalate) have no DDIs with any anticoagulant in the FDB DDI knowledgebase. This raises doubt regarding the existence of DDIs between the seven anticoagulants and these three salicylates. Indeed, upon investigating the DDIs between the three extra salicylates and these seven anticoagulants in another source [20], we discovered DDIs between the three salicylates and three of the anticoagulants. The 70 old and nine new DDIs are described in Figures 3(a) and 3(b), respectively. The reason FDB did not include these new candidate DDIs in their knowledgebase is that in these cases the drug formulation has a low potential for interaction. Nevertheless, FDB staff (JKU) confirmed that this example
demonstrates the fact that summaries of NDF-RT have the potential for supporting the discovery of new candidate DDIs. Of course, pharmacological investigation is required for each potential DDI.

Results

We derived a (non-aggregate) IAbN for the February 2015 release of NDF-RT’s Chemical Ingredients (CI) hierarchy which consists of 10,144 concepts. The complete IAbN consists of 859 ingredient groups which summarize 2,664 drug ingredients and 6,850 PP hierarchy drug concepts. We define the abstraction ratio of the IAbN to be the average number of drug ingredients per ingredient concept. The abstraction ratio of the February 2015 IAbN is 3.1 (=2,664/859). There are 813 drug ingredient concepts summarized by more than one ingredient group and each such drug ingredient is summarized by an average of 2.25 ingredient groups.

Figure 3(a). Illustration of 70 DDIs. There are 10x7=70 AVD and INL DDIs between the ten salicylates on the left and the seven anticoagulants on the right listed in Figure 3(a). Nine new candidate DDIs not appearing in FDB’s DDI knowledgebase. (b) Nine new candidate DDIs not appearing in FDB’s DDI knowledgebase. The DDIs of (a) combined with the aggregate ingredient group of salicylates in the IAbN of NDF-RT supported the discovery of nine DDIs that were confirmed by another source.

Figure 3(a). Illustration of 70 DDIs. There are 10x7=70 AVD and INL DDIs between the ten salicylates on the left and the seven anticoagulants on the right listed in Figure 3(a). Nine new candidate DDIs not appearing in FDB’s DDI knowledgebase. (b) Nine new candidate DDIs not appearing in FDB’s DDI knowledgebase. The DDIs of (a) combined with the aggregate ingredient group of salicylates in the IAbN of NDF-RT supported the discovery of nine DDIs that were confirmed by another source.

… drug concepts summarized by each ingredient group is 7.97. Figure 4 shows an excerpt of 149 of the IAbN’s ingredient groups, as the complete IAbN is too large to fit on one page. By reviewing the ingredient groups of the IAbN one can see the major types of drug ingredients used in NDF-RT’s drugs. For example, the Polymers group (Level 2: green) summarizes 26 ingredients and 81 drugs, Piperidines (Level 3: blue) summarizes 47 ingredients and 73 drugs, Tetracyclines summarizes 17 ingredients and 26 drugs, Ethanolamines summarizes 45 ingredients and 232 drugs, and Penicillins summarizes 34 ingredients and 64 drugs.

One deficiency of this complete IAbN is the relatively large number of “small” ingredient groups. For example, there are 312 (36%) ingredient groups that “summarize” only one drug ingredient. This is the reason why we developed the Aggregate IAbN methodology. The Aggregate IAbN can be used to combine these small “groups” into larger aggregate ingredient groups. We derived a few Aggregate IAbNs using bounds (b=) of 2, 6, 11, and 15, and investigated the structural properties of each resulting Aggregate IAbN (Table 1). The Aggregate IAbN in Figure 5(a), created using b=11, is composed of 30 regular ingredient groups (e.g., Ergotamines) and 88 aggregate ingredient groups (e.g., Saliycylates) which summarize 711 removed ingredient groups. Figure 5(b) illustrates how an aggregate ingredient group can be “dynamically expanded” with a software tool, so that its “summarized” removed ingredient groups can be recovered and viewed on demand.

IAbN-Based DDI Discovery Results

Table 2 shows the results of looking for new candidate DDIs using the Aggregate IAbN. Column 1 lists the interaction type. Column 2 represents the number of concepts of Group 1 (e.g., Salicylates) in NDF-RT. Column 3 represents the number of concepts in FDB among those in Column 2. Column 4 shows the number of concepts of Group 2 (e.g., Anticoagulants) interacting with Group 1 according to information in FDB’s knowledge base. Column 5 describes the number of concepts of Group 1 having no interactions with Group 2 in FDB. Column 6 describes the number of concepts among those in Column 5 for which we have found drug interactions from public sources. Column 7 lists the number of drug interactions found in those public sources.
Figure 4. An excerpt of 149 (17%) ingredient groups from the February 2015 CI hierarchy’s IABN. The smaller ingredient groups have been hidden. Child-of links are hidden for readability and longer ingredient group names are truncated. The number of ingredients and drugs summarized by each ingredient group is shown in parentheses and prepended with I: and D:, respectively. Salicylates and Aminosalicylic acids, from Figure 1(c), are highlighted in yellow.
Figure 5. (a) The Aggregate IAbN (b=11) for CI. (b) The Salicylates aggregate group expanded.

Table 1. Structural metrics for Aggregate IAbNs created using various bounds.

<table>
<thead>
<tr>
<th>Bound (b)</th>
<th># Ingredient Groups (aggregate + regular)</th>
<th># Aggregate Ingredient Groups (A)</th>
<th># Removed Ingredient Groups (R)</th>
<th>Abstraction Ratio</th>
<th>R/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>No bound</td>
<td>859</td>
<td>0</td>
<td>0</td>
<td>3.10</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>547</td>
<td>138</td>
<td>312</td>
<td>4.87</td>
<td>2.26</td>
</tr>
<tr>
<td>6</td>
<td>184</td>
<td>102</td>
<td>675</td>
<td>14.5</td>
<td>6.61</td>
</tr>
<tr>
<td>11</td>
<td>88</td>
<td>58</td>
<td>771</td>
<td>30.3</td>
<td>13.29</td>
</tr>
<tr>
<td>15</td>
<td>45</td>
<td>37</td>
<td>814</td>
<td>59.2</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 2. DDI discovery results

<table>
<thead>
<tr>
<th>Column 1: Interaction Type</th>
<th>Col 2: #concepts of Group 1 in NDF-RT</th>
<th>Col 3: #concepts of Group 1 in FDB</th>
<th>Col 4: #concepts of Group 2 in FDB w/ no DDI in FDB</th>
<th>Col 5: #Group 1 Concepts w/ new DDI candidates</th>
<th>C6: #Col 5 concepts with new DDI candidates</th>
<th>Col 7: #new DDI candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salicylates/Anticoagulants</td>
<td>18</td>
<td>13</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Salicylates/Heparin</td>
<td>18</td>
<td>13</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Salicylates/Unicosurics</td>
<td>18</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Salicylates/Antidiabetics, oral</td>
<td>18</td>
<td>13</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>Salicylates/Valproic acid</td>
<td>18</td>
<td>13</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Hydantoin/Selected anticoagulants</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Hydantoin/isoniazid</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hydantoin/ Folic acid; Pyrimethamine</td>
<td>8</td>
<td>4</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Hydantoin/Sulfonamides</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Hydantoin/Cimetidine; Ranitidine</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 2 lists five new candidate ingredient pairs (rows) for Salicylates. For each such pair Column 5 lists the number of salicylates from NDF-RT appearing in FDB but not having such a DDI. From those, Column 6 lists the number of salicylates that have DDIs with drugs of Group 2 in a public source and Column 7 lists the number of such DDIs. For example, row 1 reports the case described in Figure 3. However, the three extra salicylates e.g. mesalamine have no interactions with dicumarol, warfarin, and anisindione in FDB because their drug formulations produce localized concentrations in the gastrointestinal tract but not systemic blood levels where they will interact with anticoagulants.

In the last five rows of Table 2, we list ingredient pairs with hydantoins. Looking for hydantoins in NDF-RT that also appear in FDB but have no DDIs, a base formulation representation of the fosphenytoin ingredient in NDF-RT is found, for which public sources list DDIs. A review of FDB’s knowledge base by a coauthor (JKU) found the DDIs associated with the salt formulation representation of this ingredient, fosphenytoin sodium, rather than with fosphenytoin. In this case, the candidate base ingredient does not exist in pharmaceutical formulations and thus there is no clinical data gap. However, there may be other instances, e.g., erythromycin base with physical dosage forms that would be considered candidates for inclusion in DDIs. Hence, these DDI candidates are legitimate.

Discussion

The development of the IAbN (Ingredient Abstraction Network) represents an important first step in the summarization of NDF-RT’s drug concepts according to their various classifications. The IAbN makes it possible to compactly visualize the major types of ingredients, and the drugs which contain them, as they exist in NDF-RT. One reason that the IAbN derivation methodology works well for the CI hierarchy is that the majority of drug ingredient concepts in NDF-RT are leaves (i.e., have no children) or are near the bottom (leaves) of the CI hierarchy.

The IAbN derivation approach is applicable beyond the CI hierarchy. For example, it is possible to apply the IAbN derivation methodology to the Mechanism of Action (MOA) hierarchy, summarizing NDF-RT’s drugs according to their mechanisms of actions, rather than their chemical ingredients. In a future study, we will investigate the structural properties of IAbNs derived for NDF-RT’s other classification hierarchies.

The Aggregate IAbN provides a secondary summarization mechanism that allows control over the granularity of summarization provided by an IAbN. The Aggregate IAbN lets a user see a very compact representation of a terminology hierarchy. Fine control is possible by choosing different values for b (the bound). The software tool (under development) for displaying IAbNs lets the user “re-expand” aggregate ingredient groups and lets her/him inspect previously hidden removed ingredient groups. In other words, the knowledge in the removed ingredient groups is not lost; it is just hidden and can be displayed by the tool upon demand. One issue to consider in future research is when one should use an Aggregate IAbN versus the complete IAbN. A usability study of the software tool and a study of the trade-offs between different abstraction ratios in Aggregate IAbNs will be performed.

In future research, we will investigate the use of alternate methods for controlling the amount of information of an IAbN that is displayed. This will include generating IAbNs that are created, e.g., by choosing a specific ingredient group and then viewing all of its ancestor and/or descendant ingredient groups. Additionally, we will investigate the creation of Aggregate IAbNs according to the number of PP drug concepts, rather than the number of ingredient concepts. This approach would provide an alternate secondary summary that highlights which ingredient groups summarize more PP drugs. Different users with various professional profiles may prefer different options according to their emphasis on ingredients or on drugs.

One significant difference between the IAbN and our previously developed abstraction networks is that not every concept in the CI hierarchy is summarized by an ingredient group. In our previous AbNs, every concept was always summarized by at least one AbN node [2]. However, since a CI hierarchy concept may be neither a classification ingredient nor a drug ingredient, it may be omitted from the IAbN, e.g., Ethyl biscoumacetate in Figure 1. In fact, the majority (6,621, 65.3%) of concepts in the CI hierarchy are not summarized by any ingredient group. This situation occurs for several reasons. The CI hierarchy was primarily imported from MeSH and many of the concepts from MeSH are too general and do not represent ingredients that could be used in drugs. Other ingredient concepts may not be relevant to the drugs in NDF-RT, as no drug includes them as an ingredient.

The goal of the IAbN is to summarize NDF-RT’s drugs according to their ingredients. Thus, it is not necessary to summarize CI concepts that are not used as ingredients in drugs. However, it is conceivable that there may be applications which need access to these concepts. In a future study we will investigate ways of summarizing non-classification/non-drug ingredient concepts. One potential idea is to associate each such concept with its closest classification ingredient. Using such an approach, Ethyl biscoumacetate would be summarized by the 4-Hydroxycoumarins ingredient group.
Conclusions

In this paper, we introduced the Ingredient Abstraction Network (IAbN) to summarize the concepts in NDF-RT’s Chemical Ingredients hierarchy. A parametric method for controlling the level of summarization in a secondary abstraction network, called Aggregate IAbN, was introduced. The IAbN was shown to support the discovery of new candidate drug-drug interactions in a DDI knowledgebase.

Acknowledgements

We thank Michael Lincoln of the VHA and Mark Erlbaum of Apelon for sharing their insights into NDF-RT and its history with us. This work was partially supported by NIH grant 1R01CA190779-01.

References

Machine Learning Approaches for Detecting Diabetic Retinopathy from Clinical and Public Health Records

Omolola Ogunyemi, PhD,1,3,4 Dulcie Kermah, MPH,2,3
1Center for Biomedical Informatics, 2Biostatistics Core, 3Charles Drew University of Medicine and Science, Los Angeles, California
4University of California Los Angeles, Los Angeles, California

Abstract

Introduction: Annual eye examinations are recommended for diabetic patients in order to detect diabetic retinopathy and other eye conditions that arise from diabetes. Medically underserved urban communities in the US have annual screening rates that are much lower than the national average and could benefit from informatics approaches to identify unscreened patients most at risk of developing retinopathy.

Methods: Using clinical data from urban safety net clinics as well as public health data from the CDC’s National Health and Nutrition Examination Survey, we examined different machine learning approaches for predicting retinopathy from clinical or public health data. All datasets utilized exhibited a class imbalance.

Results: Classifiers learned on the clinical data were modestly predictive of retinopathy with the best model having an AUC of 0.72, sensitivity of 69.2% and specificity of 55.9%. Classifiers learned on public health data were not predictive of retinopathy.

Discussion: Successful approaches to detecting latent retinopathy using machine learning could help safety net and other clinics identify unscreened patients who are most at risk of developing retinopathy and the use of ensemble classifiers on clinical data shows promise for this purpose.

Introduction

Diabetic retinopathy arises when excess glucose in the bloodstream resulting from diabetes mellitus causes damage to the blood vessels of the retina. Among US adults between the ages of 20 and 74 years, diabetic retinopathy is the leading cause of blindness.1,2 Diabetes affects an estimated 29.1 million people in the United States.3 In a Centers for Disease Control and Prevention (CDC) assessment performed between 2005 and 2008, 4.2 million or 28.5% of people with diabetes aged 40 years or older in that time period had diabetic retinopathy.3

Annual eye examinations for diabetic patients are recommended in order to detect and treat diabetic retinopathy in a timely manner, since blindness from this condition is preventable with early detection and the use of laser photocoagulation therapy. While the US national annual eye screening average for diabetic patients is 60%,4-8 some studies of the urban safety net setting have shown annual screening rates for inner-city diabetic patients to be lower than 25%.9-11 Previously, we presented findings from a study that sought to examine issues affecting that disparity between national and urban safety net screening rates by highlighting the feasibility and challenges of implementing teleretinal screening for diabetic retinopathy in an urban safety net setting facing eyecare specialist shortages.12-15

We also examined the potential for developing predictive models for detecting diabetic retinopathy on safety net clinic data. Here, we build on knowledge gained from our earlier diabetic retinopathy predictive modeling work;16 and we also examine the utility of using National Health and Nutrition Examination Survey (NHANES) public health data collected by the CDC in our model development effort, inspired by a study from South Korea that utilized similar public health data from that country to create predictive models for diabetic retinopathy.17 Our long-term goal is to develop predictive models for diabetic retinopathy that are appropriate for the safety net setting.

Risk factors for diabetic retinopathy cited in the literature include duration of diabetes,18-20 high blood glucose/poor blood sugar control,18-21 high blood pressure,18-21 dyslipidemia,18 high cholesterol,19 pregnancy,18 nephropathy,20 and obesity.18 Other known risk factors for diabetic retinopathy include inflammation,18 puberty (in Type 1 diabetes),18 ethnicity,18 insulin treatment (for Type 2 diabetes - related to poor blood glucose control), tumor necrosis factor receptors,22 and smoking23 (in one study, pack-years of smoking were found to be borderline significant in predicting retinopathy in younger adults;24 however, other studies have found the relationship between retinopathy and smoking to be inconsistent25, 26). Risk factors such as high blood glucose, duration of diabetes and high blood pressure are considered to be stronger predictors of retinopathy than other risk factors.18 Models that
incorporate some or all of these known risk factors as well as other factors that may not have been considered in the literature could be useful for prediction. Aside from tumor necrosis factor receptors, most of the risk factors listed above are routinely collected and stored in electronic health records in the course of care for a diabetic patient and thus could be used to create predictive models to screen for latent retinopathy in diabetic patients who have not received an annual eye examination as recommended in clinical practice guidelines.

There are different stages of diabetic retinopathy, which in order of increasing severity are: mild non-proliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, and proliferative diabetic retinopathy. On eye examinations, there are visible developments in the retina as diabetic retinopathy progresses that aid clinicians in distinguishing one stage of retinopathy from another. These include micro-aneurysms, intra-retinal hemorrhages, retinal ischemia (cotton-wool spots), venous beading, and finally, the proliferation or growth of fragile new blood vessels that can bleed on the retina’s inner surface.27 For the present study, our goal is to develop methods to identify patients at high risk of diabetic retinopathy in order to have them come in to a care site for a teleretinal eye screening or an in-person eye examination. This means that the methods developed will predict the presence or absence of retinopathy generally; retinopathy staging is not attempted (to accurately stage retinopathy, data from digital retinal images would be required).

For the CDC NHANES dataset utilized in the present study, which contains merged data from 2005-2006 and from 2007-2008, we found that only 13% (158 instances) involved patients who had diabetic retinopathy, while 87% (1081 instances) did not. In our earlier predictive modeling work, the dataset collected from urban safety net clinics corresponded to 513 diabetic patients, with approximately 25% of instances (130 instances) having an outcome of diabetic retinopathy and roughly 75% of the instances (383 instances) having an outcome of no diabetic retinopathy. The diabetic retinopathy rate of 25% in that clinical dataset is roughly in line with the projected US rate of 28.5% between 2005 and 2008 referenced above. Using standard classifiers, the best classification result on the clinical data from 513 patients was achieved with a Bayesian network that had a sensitivity of 26.2%, a specificity of 94.5%, an Area Under the ROC Curve (AUC) of 0.71, a negative predictive value of 79%, and a positive predictive value (precision) of 61.8%.16

One of the difficulties with predicting diabetic retinopathy solely from clinical or public health data for screening outreach purposes (without the benefit of digital retinal imaging features that are clearly only associated with retinopathy) is that there is some overlap between the classes corresponding to “retinopathy” and “no retinopathy,” which makes the two classes difficult to separate. Additionally, as described above, datasets based on clinical or public health records may be mostly skewed towards “no retinopathy.” A class imbalance is said to occur when one of the outcomes of a learning problem, especially the outcome of interest, is underrepresented in the dataset from which predictive models are to be learned.28 As each dataset used for the present study exhibits a class imbalance and we have previously had modest results on the clinical dataset with standard classifiers, we turn to machine learning approaches designed to deal with the class imbalance problem in this study. Galar et al identify four key approaches to dealing with the class imbalance problem: (1) techniques that modify existing algorithms for a standard classifier in order to emphasize the significance of satisfactorily classifying the minority class; (2) data preprocessing methods, such as undersampling the majority class or oversampling the minority class; (3) cost sensitive approaches that combine both algorithm and data preprocessing methods, and, (4) ensembles of weak learners (classifiers) that take advantage of data preprocessing and other methods.28 For this study, we focused on the use of ensembles of weak learners that take advantage of data preprocessing methods.

Methods

Approval to use clinical data for the study was obtained from the Charles Drew University Institutional Review Board.

Data sources:
Clinical data for the study were previously obtained from six federally qualified health centers that serve as primary care clinics for un- and under-insured patients in South Los Angeles. This was done through a retrospective review of medical records for 513 patients with type 2 diabetes who received an eye examination for diabetic retinopathy from ophthalmologists via teleretinal screening and who obtained care at the clinical sites in 2011.
Public health data for the study was obtained from the US NHANES cross-sectional data set. The National Health and Nutrition Examination Survey is conducted by the National Center for Health Statistics, using a stratified multistage probability design to obtain a representative sample of the total civilian, non-institutionalized US population. Since 1999, the NHANES has released data at 2-year intervals. The NHANES collects questionnaire data during face-to-face home interviews and includes a physical examination, as well as the collection of laboratory data. We used data from 2005-2006 and 2007-2008. These years had the only datasets that matched 95% of the variables utilized in the previously mentioned Korean study.\textsuperscript{17}

Sample from NHANES:
There were a total of 20,497 participants from NHANES 2005 to 2008. For this analysis, we only included persons with diabetes (N=2,874). Of those who had diabetes, participants who had complete data on the presence or absence of diabetic retinopathy were included in the analysis. Our analytical sample was based on a total of 1239 people with diabetes who had been told whether or not they had retinopathy. Participants were identified as being diabetic if they met at least one of the following criteria; plasma fasting glucose ≥ 126 mg/dL, serum glucose ≥ 200mg/dL, glycohemoglobin ≥ 6.5%, responding “Yes” to the question “Other than during pregnancy, have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?,” responding “Yes” to the question, “Are you now taking insulin?,” responding to the question “Are you now taking diabetic pills to lower your blood sugar? These are sometimes called oral agents or oral hypoglycemic agents.”

Classification Methods:
Variables with 50% or more of their values missing were not included in the datasets used for machine learning. Missing data for the remaining variables (less than 50% of values missing) were handled by using imputation techniques. For the datasets analyzed, we performed feature subset selection using the Lasso.\textsuperscript{29} Since we used standard (single) classifiers in our previously published study, for this study, we use ensembles, which combine classifiers and may perform better than a single classifier approach. We learned ensemble classifiers based on decision tree learners designed to handle class imbalances such as RUSBoost,\textsuperscript{30} which utilizes majority class undersampling. For contrast, we also learned ensemble classifiers using AdaBoost.M1,\textsuperscript{31} which uses adaptive boosting to combine the weighted output of several weak learners to produce a boosted classification output. On its own, AdaBoost.M1 has no special accommodation for class imbalances. The ensemble classifiers were learned on the full feature set as well as the feature subsets obtained. We reserved 20% of each dataset for testing and then performed 10-fold cross validation on the remaining 80% of the dataset, selecting the best classifier from the cross-validation process for use on the reserved test set. Analyses were performed using MATLAB\textsuperscript{32} and Weka.\textsuperscript{33}

For each classifier, we measured sensitivity or the true positive rate (the total number of cases classified as having diabetic retinopathy divided by the total number of cases actually involving retinopathy), specificity or the true negative rate, the AUC, which represents the trade-off between the true positive rate/sensitivity and the false positive rate or 1 – specificity, and accuracy (the total number of correctly classified cases divided by the total number of cases).

Results

Table 1: Variables gathered from clinic that potentially impact development of diabetic retinopathy

<table>
<thead>
<tr>
<th>Clinical variables that might impact diabetic retinopathy risk</th>
<th>Gender</th>
<th>Age</th>
<th>Ethnicity/race</th>
<th>Education</th>
<th>Insulin dependence</th>
<th>Number of years patient has had diabetes</th>
<th>Hemoglobin A1C value</th>
<th>Co-morbid conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Gender</td>
<td>Age</td>
<td>Ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Ethnicity/race</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Insulin dependence</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Number of years patient has had diabetes</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Hemoglobin A1C value</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Co-morbid conditions</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Hypertension</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Nephropathy</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Depression</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Other (hypothyroidism, etc.)</td>
<td></td>
<td></td>
<td>ethnicity/race</td>
<td>Education</td>
<td>Insulin dependence</td>
<td>Number of years patient has had diabetes</td>
<td>Hemoglobin A1C value</td>
<td>Co-morbid conditions</td>
</tr>
</tbody>
</table>

985
The complete set of predictors/features utilized for the clinical data is presented in Table 1. Table 2 shows the feature subset obtained after applying the Lasso to the dataset in Table 1. Table 3 shows the results of 10-fold cross validation on the 80% of the clinical dataset set aside for this purpose, first using the entire set of features and next, using the feature subset. It also shows the results of applying the best ensemble model obtained from 10-fold cross-validation to the 20% of cases set aside solely for testing.

Table 2: Subset of clinical variables following feature-subset selection using the Lasso

<table>
<thead>
<tr>
<th>Clinical variables selected from feature subset selection using the Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Marital Status</td>
</tr>
<tr>
<td>Hemoglobin A1C value</td>
</tr>
<tr>
<td>Number of years patient has had diabetes</td>
</tr>
<tr>
<td>Co-morbid conditions</td>
</tr>
<tr>
<td>Hypertension</td>
</tr>
<tr>
<td>Previous diagnosis &amp; treatment of retinopathy</td>
</tr>
</tbody>
</table>

Table 3: Results for ensemble classifiers on clinical data

<table>
<thead>
<tr>
<th>Averaged results for ensemble classifiers following 10-fold cross validation using all features</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUS Boost Ensemble Average</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>AUC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results for best ensemble on 20% set-aside test set using all features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best RUSBoost Ensemble</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>AUC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Averaged results for ensembles following 10-fold cross validation using feature subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUSBoost Ensemble Average</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>AUC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results for best ensemble on 20% set-aside test set using feature subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best RUSBoost Ensemble</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>AUC</td>
</tr>
</tbody>
</table>

The complete set of predictors/features utilized for the public health data is presented in Table 4. Table 5 shows the feature subset obtained after applying the Lasso to the dataset in Table 4. Table 6 shows the results of 10-fold cross validation on the 80% of the clinical dataset set aside for this purpose, first using the entire set of features and next, using the feature subset. Table 6 also shows the results of applying the best ensemble model obtained from 10-fold cross-validation to the 20% of cases set aside solely for testing.
### Table 4: Variables gathered from the NHANES 2005-2008 dataset

<table>
<thead>
<tr>
<th>Public health (NHANES) variables that might impact diabetic retinopathy risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Sex</td>
</tr>
<tr>
<td>Smoking status</td>
</tr>
<tr>
<td>Alcohol consumption</td>
</tr>
<tr>
<td>Insulin therapy</td>
</tr>
<tr>
<td>On diabetes pills</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
</tr>
<tr>
<td>Diagnosed hypertension</td>
</tr>
<tr>
<td>Diagnosed diabetic nephropathy</td>
</tr>
<tr>
<td>Body mass index</td>
</tr>
<tr>
<td>hemoglobin A1C</td>
</tr>
</tbody>
</table>

### Table 5: Subset of NHANES variables following feature-subset selection using the Lasso

<table>
<thead>
<tr>
<th>NHANES variables selected from feature-subset selection using the Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Hemoglobin</td>
</tr>
<tr>
<td>On non-drug diabetes interventions</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
</tr>
<tr>
<td>Education</td>
</tr>
</tbody>
</table>

### Table 6: Results for ensemble classifiers on public health (NHANES) data

<table>
<thead>
<tr>
<th>Averaged results for ensemble classifiers following 10-fold cross validation using all features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>AUC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results for best ensemble on 20% set-aside test set using all features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>AUC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Averaged results for ensembles following 10-fold cross validation using feature subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>AUC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results for best ensemble on 20% set-aside test set using feature subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Sensitivity</td>
</tr>
<tr>
<td>Specificity</td>
</tr>
<tr>
<td>AUC</td>
</tr>
</tbody>
</table>
Discussion

The results show that the clinical dataset was moderately predictive for diabetic retinopathy, with the best RUSBoost ensemble having an accuracy of 73.5%, sensitivity of 69.2%, specificity of 55.9%, and AUC of 0.72 on previously unseen instances (the test data set aside). As the AdaBoost ensemble results show by comparison, there is a trade off of an increase in sensitivity with majority class undersampling in RUSBoost for a decrease in specificity. For the purpose of identifying individuals who have not yet received an annual eye examination but may have latent retinopathy, the increase in sensitivity provided by RUSBoost is important. In previous work on the same dataset, the best standard classifiers achieved stellar specificity (>90%) but poor sensitivity (<28%), which made them less suitable for screening purposes. The best AdaBoost ensemble achieved better sensitivity than a standard Bayesian network classifier (34.6% versus 26.2%) but much worse specificity (63.7% versus 94.5%). A greater improvement in sensitivity and reduction in specificity was also seen for the best RUSBoost ensemble results (sensitivity of 69.2% versus 26.2% for the Bayesian network classifier and specificity of 55.9% versus 94.5% for the Bayesian network classifier). The improved sensitivity results on the clinical data with RUSBoost are encouraging, since the data include features that are routinely collected by and available to clinics treating diabetic patients and could provide the basis of targeted outreach to noncompliant patients.

The classifier ensembles developed on the public health data were not useful, with AUC’s indicating that they were no better than flipping a coin at discriminating between cases of retinopathy and cases of no retinopathy. While the accuracy results for AdaBoost were high on the public health dataset, it is clear from the best AdaBoost sensitivity results (3.2%) and the best specificity results (85.8%) that the ensembles tended towards predicting “no retinopathy” in almost every situation and was not as adaptive in adjusting to misclassifications as might have been expected.

The NHANES dataset utilized relies on physical examinations as well as patient questionnaires for some key variables; poor responses to the questionnaires may have led to poorer quality data for learning. For example, duration of diabetes, which is known to be a major risk factor for retinopathy had to be derived from questionnaire responses about the age at which a patient was diagnosed with diabetes (subtracting the age of diagnosis from the current age). There were negative durations in some cases and over 70% missing data for this key variable; these inaccuracies meant that we were not able to include duration of diabetes for the public health data analyses in the present study. It is likely that a public health dataset with more variables related to retinopathy measured clinically and fewer missing data would produce better results. However, a recently published study that looked only at a subset of 266 people over the age of forty from NHANES who were newly identified by the paper’s authors as having diabetes had more promising results with regards to predicting retinopathy, with a best AUC of 0.74, a low positive predictive value (precision) of 22% and a high negative predictive value of 99%. 34

Future work includes collecting additional clinic data, applying novel methods for learning with a class imbalance to the data and developing software from the predictive models developed. The software is intended for clinics to use for targeted outreach to patients who have not received an annual eye examination (in addition to their usual outreach to diabetic patients). Predictive models that are able to accurately identify individuals with diabetes who may have developed retinopathy but are not yet aware of it would be invaluable for outreach purposes in order to prevent avoidable blindness from diabetic retinopathy.

Acknowledgments

This project was supported by the NIH under grant numbers U54 MD007598, U54 RR026138-01S2, and S21MD000103.

References


988
Understanding patients’ health and technology attitudes for tailoring self-management interventions

Katie O’Leary¹, Lisa Vizer, PhD¹, Jordan Eschler¹, James Ralston, MD, MPH², and Wanda Pratt, PhD¹
¹University of Washington, Seattle, WA; ²Group Health Research Institute, Seattle, WA

Abstract
Healthcare providers are moving towards tailoring self-management interventions to include the communication technologies patients use in daily life. Accurate understanding of patients’ attitudes towards both technology and involvement in managing chronic conditions will be critical for informing effective self-management strategies. The tailoring of these interventions, however, could be undermined by providers’ implicit biases based on patient age, race, and education level that have been shown to negatively affect care. To inform the design and tailoring of self-management interventions, we elicited attitudes toward technology use and participation in care of 40 participants in a maximum variation sample. The analysis revealed three participant clusters—“Proactive Techies,” “Indie Self-Managers,” and “Remind Me! Non-Techies”—that represent varying attitudes toward health behaviors and technologies that were independent of race, education level, and age. Our approach provides insight into how people prioritize important values related to health participation and technology.

Introduction
Self-management of chronic conditions, such as diabetes and asthma, creates considerable burden for patients. Effective self-management requires patients to be active participants in treatment by taking medications, making lifestyle changes, and regulating emotions¹,². The day-to-day skills involved in self-management include identifying problems, selecting solutions, using information sources, collaborating with providers, changing behavior, and evaluating results². This dynamic process goes well beyond adhering to treatment plans, and considers the patient holistically in the context of the attitudes and values that influence their everyday behaviors.

Self-management interventions are designed to alleviate some of the burden of chronic illness care. Self-management support through the internet, over the phone, and in person have proven effective, and the delivery of programs via SMS and other mobile platforms are promising³–⁸. However, tailoring and personalization of self-management interventions is limited by the lack of methods for understanding how patients prioritize health information and tools to accomplish self-care goals. Moreover, implicit biases of providers, such as ascribing preferences for health information and treatments on the basis of age, race, literacy, and gender, can differentially affect provision of chronic care for underrepresented minorities⁹–¹⁶. Therefore, understanding patients’ attitudes toward health and technology is especially important for informing unbiased and patient-centered interventions that augment patients’ health self-management strategies in everyday life.

To address the lack of understanding of the attitudinal factors that influence patients’ preferences for health information and technology, we studied how patients prioritize tools (e.g. email, cell phones, paper tools) to strengthen their accountability for health self-management. We found that patients’ attitudes toward mobile technology, accountability, and goal-setting were the most distinguishing and influential attitudes on patients’ approaches to health self-management. We also found that patients’ attitudes and values were more influential than demographic characteristics—especially race, age, and education level—on defining health self-management behaviors. We contribute a description of a method for eliciting preferences for self-management support, including attitudes toward adopting technology to attain health goals. Our empirical evidence from this study can inform new approaches to clustering patients and tailoring appropriate self-care support based on the tools, information, and goals they perceive to be essential to managing their health.

Related Work
Previous research on self-management, disparities in use of e-health services, and patterns of technology adoption inform our work. Below, we briefly summarize this literature and describe how our approach to clustering patients based on health and technology attitudes contributes new insights for supporting self-management.
Patient Self-management

Studies show that self-management interventions are effective in supporting patients to manage chronic illness\(^1\)\(^7\)\(^-\)\(^9\). In particular, interventions that systematically support patients to build skills in the three major domains of self-management—medical, role, and emotional—offer the most comprehensive support for self-care\(^1\)\(^2\). Within all three domains, self-management interventions are more likely to succeed if they target problems and strategies that patients perceive as most significant to their health, because the focus on patients’ strategies and concerns is motivating, engaging, and empowering\(^2\). Our approach provides researchers with a holistic perspective well-suited to tailoring self-management programs based on what patients perceive as essential to managing their health. As many healthcare providers are beginning to engage patients in care activities online and over mobile devices, we investigated both health and technology attitudes to understand technology adoption in support of patients’ overarching health goals.

Intervention research

Research on targeting self-management interventions has focused on understanding the sources of variation in the use of e-health services. For example, Goel et al\(^20\) investigated racial disparities in e-health services enrollment, and found that barriers to internet access did not explain these disparities; rather attitudinal barriers were significant. Similarly, Clark et al\(^21\) found that understanding attitudes and health priorities is particularly important for tailoring self-management interventions for socioeconomically vulnerable older adults. Finally, Lyles et al\(^22\) found that differences in portal use by race and ethnicity were not fully explained by differences in age, sex, sociodemographics, health status, or provider factors—particularly for black patients. Therefore, demographic characteristics provide limited help in understanding differences in portal use and how to tailor self-management support programs to the technologies patients use. Our approach to investigating values that influence patients’ use of e-health services can provide alternative perspectives on the attitudes that drive patient engagement in self-management programs.

The importance of understanding the attitudes and beliefs that influence engagement with health technologies may become even more important as intelligent information systems are deployed to support self-management. For example, recent approaches to mHealth care include intelligent systems that adapt to users depending on the patient’s beliefs about illness, medication, and information needs\(^23\). Researchers like Clark et al\(^21\) and Koch et al\(^24\) have called for a greater emphasis on the “self” in self-management, and have brought attention to the importance of understanding patient’s priorities, expectations, and expertise in the care process. Moreover, Valdez et al\(^25\) recently emphasized a “patient work approach” to designing applications aligned with patients’ and their family members' health-related activities. As mHealth service design becomes more adaptive and personalizable, the “self” in self-management will become a focal point for tailored interventions. We demonstrate an approach for “getting to know” the patient that provides insight into how patients prioritize different values related to health and technology.

Finally, Pew\(^26\) data suggests that people from underserved populations, such as people with low education and ethnic minorities, are highly engaged with technology, even compared to privileged populations. These findings undermine implicit assumptions about correlations between low education level or racial minorities and low technology use. However, research by Sarkar et al\(^27\) also suggests that these populations are less likely to use technology to manage their health, which points to the need for researching the disparity between technology adoption for everyday life compared to health management. We demonstrate that qualitative research may provide important insights into the social and attitudinal barriers to adopting e-health services that can explain low technology adoption for health.

Methods

We used a mixed-methods approach that combined in-depth, semi-structured interviews with 40 participants, using a card sorting procedure called Q-methodology\(^28\)\(^,\)\(^29\) to elicit attitudes and opinions from persons managing a chronic illness. Q-methodology has been used in health\(^30\),\(^31\) and technology\(^32\) settings to understand patient perspectives. We conducted the interviews and Q-methodology procedure in participants’ homes, usually at their kitchen tables. Each interview lasted 1.5 to 2 hours and focused on understanding the participants’ health goals, priorities, and attitudes. The interviews were transcribed, and identifiable information was redacted to protect participants’ confidentiality. The study was reviewed and approved by the Internal Review Board of the Group Health Research Institute.
Participants
We recruited older adults with diabetes Type 1 and Type 2 (n=20), and mothers of children with asthma (n=20). We recruited these two distinct samples to maximize the variation in attitudes toward health self-management and technology use. We sampled mothers of children with asthma (rather than fathers or both parents) because women in this age group have ongoing health care needs relevant to reminders, such as screening for cervical cancer. All participants (see Table 1) were sampled from Group Health Cooperative, a large integrated healthcare delivery system in Washington State. We purposely oversampled racial and ethnic minorities, who would substantially benefit from improvements in care related to technology and chronic illness management. These interviews were the second round of interviews for 34 of the 40 participants. We had recruited these 34 participants by mail and phone from a bank of 586 people for this three-year multi-part study. The remaining 6 people were recruited from a new bank of 99 people (25 older patients, and 74 mothers). Older adults (10 of 20 were male) had a median age of 73 years, and mothers had a median age of 38 years. After the interview, they received $50 for completing the study.

In-depth Interviews and Q-methodology
The interview protocol contained open-ended questions about health goals, and a reflection on the challenges and personal significance of nationally recommended health tasks appropriate to the age and diagnoses of each participant. The Q-methodology procedure was conducted at the end of the interview. The data collection procedure for Q-methodology consists of a set of subjective statements printed on cards that the participant arranges according to what is most personally significant to them. It is a highly structured data collection procedure wherein participants arrange the statements in a grid that is shaped in a normal distribution. The extremes of the distribution represent the statements that the participant feels most strongly about (See Figure 1): “Most Agree” to “Most Disagree” are located at either end of the Q grid, with “Neutral” in the middle. The statements that participants place at the extremes receive the most weight in the cluster analysis of the data, revealing the health and technology attitudes that distinguish unique patient perspectives (See Table 2).

Our Q statement set consisted of 27 statements based on a health-sciences theoretical framework. The statements were categorized into two groups: (1) attitudes toward health self-management, especially related to core self-management skills; and (2) attitudes toward using and adopting technology. We developed statements in the first category on self-management attitudes based on the three large domains of work in chronic illness care originally identified by Corbin and Strauss\(^1\) and incorporated into self-management support programs by Lorig and Holman\(^2\). These included the work of medical management, emotional support, and redefining life roles. We used prior instruments of patient activation to inform statements representing patients’ attitudes towards self-management in these domains. These include the Patient Activation Measure (PAM)\(^3\) and the Patient Assessment of Chronic Illness Care (PACIC)\(^5\). The PAM assesses patients’ self-management beliefs, knowledge, skills, and confidence. The PACIC scale assesses the extent to which the patient has received care congruent with the Chronic Care Model and includes a subscale of patient activation. We used both of these scales to inform a set of statements on a continuum of patients’ self-management beliefs, confidence, and practices\(^6\). These statements allowed us to observe how patients prioritize health and technology behaviors for health self-management. We derived the wording of statements from qualitative data from a previous round of interviews with these participants\(^\_3\).
Analysis
We analyzed the data with PQMethod\textsuperscript{38}, a widely used application for conducting exploratory cluster analyses of Q-methodology data. We conducted a principal components analysis that extracted 8 components, 3 of which explained 53% of the variance in the data. We chose to retain only 3 components for interpretation based on the standard criteria for factor rotation (i.e. proportion of variance, scree test, and interpretability)\textsuperscript{39,40}. We rotated the three components using varimax, a standard orthogonal rotation method.

We used the distinguishing statements—statements that distinguish each cluster based on statistical significance—to inform the qualitative analysis of the interview data (See Table 2). By checking the themes that emerged from the distinguishing statements in the cluster analysis against the qualitative data, we were able to interpret the clusters in the context of each participant’s health goals, and to report on their attitudes and beliefs using their own words. Moreover, this method of qualitative validation is especially important when working with marginalized populations\textsuperscript{41}. Participant quotes are incorporated into our results, with A\# identifiers denoting the words of mothers with children who have asthma, and D\# identifiers denoting quotes from diabetes patients.

Results
The principle components analysis resulted in three clusters that represent three unique perspectives on both health attitudes and tools that are essential to health self-management. These three components yielded excellent interpretability and a “simple structure,” with only 4 participants whose views are excluded from analysis due to their loading significantly onto more than one component. Clusters 1 (n=14), 2 (n=13), and 3 (n=9), do not correlate with age, race, educational background, portal use, gender, or disease. Rather, each cluster represents a unique attitude towards health self-management held by a diverse subset of the sample (see Table 1). We describe each cluster below, and characterize them as: (1) Proactive Techies; (2) Indie Self-Managers; and (3) Remind Me! Non-Techies. We denote the distinguishing statements for each cluster with the ranking assigned to it, on a scale of Most Agree (+3) to Most Disagree (-3), by participants of the cluster.

Table 1. Participant demographics by attitude cluster and cohort. A=asthma cohort; D=diabetes cohort.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Proactive Techies</th>
<th>Indie Self-Managers</th>
<th>Remind Me! Non-Techies</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>Participants</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Hispanic</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Asian</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>White</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Mixed/Other</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School/GED</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Some College</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4-year degree</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>More than 4-year</td>
<td>1</td>
<td>-</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Portal User</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 2. Example distinguishing statements for each attitude cluster, (p>0.05). +3=Most Agree; -3=Most Disagree.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Distinguishing Statement</th>
<th>Statement Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proactive Techies</strong></td>
<td>I like using electronic communication with my doctor.</td>
<td>+3</td>
</tr>
<tr>
<td></td>
<td>I’m more likely to follow through on paper mail reminders.</td>
<td>-3</td>
</tr>
<tr>
<td><strong>Indie Self-Managers</strong></td>
<td>I’m the one who is responsible for taking care of my health.</td>
<td>+3</td>
</tr>
<tr>
<td></td>
<td>My health will improve if I get better at mobile technology.</td>
<td>-3</td>
</tr>
<tr>
<td><strong>Remind Me! Non-techies</strong></td>
<td>My health will improve if I set health goals with my doctor.</td>
<td>+3</td>
</tr>
<tr>
<td></td>
<td>I mostly keep track of health tasks in my head.</td>
<td>-3</td>
</tr>
</tbody>
</table>

Cluster 1: “Proactive Techies”
Fourteen participants (3 male) define this cluster, including six participants from the diabetes cohort (ages 63-74) and eight from the asthma cohort (ages 32-49) (see Table 1). This cluster has a high technology profile. The participants who define this cluster strongly agree that cell phones (+2), electronic communication (+3), and knowledge of the science behind health advice (+2) are essential to health self-management. They somewhat agree that electronic reminders will reduce the stress of managing their health (+1) and that online calendars are the best way to manage their health (+1). Consistent with their preference for electronic methods, they strongly disagree that paper calendars (-3) and reminders (-3) are useful for health self-management.

Patients in this cluster were distinguished by their emphasis on the connection between health self-management and technology use. These patients used technology to take the initiative with their health care by proactively tracking health indicators, understanding scientific health information, and setting personal goals. This health tracking behavior supported these patients to set health goals for themselves and to take responsibility for their health outcomes. In particular, older adults with diabetes described using technology to track self-care activities. For example, D15, a black man with some college education, recounted how he used a spreadsheet to track his blood glucose levels, and sent it in an email to give his doctor a “heads up”:

“It's going back about 18 months ago, I went and got it [blood glucose] on Excel in a nice little graph that looked good to me ... when we had my primary care doctor and we met, she was able to say okay, you need to up your medication...[the doctor] had a heads up before I got there, not only from her examination but from me sending her information.”

Similarly, D6, a black woman with a 4-year degree, described how she used the patient portal to track a detailed history of her A1C test results:

“[On the patient portal] there's information about the test, why they take the test, what the numbers are, whatever it is. So you know where you should be, and then they keep it - I think mine started 2008, they keep it for a long time and they have a chart so you can see where your ups and downs are.”

D7, another black woman with a 4-year degree, described how she uses her cell phone for keeping track of her health, “I keep my appointments in there [cell phone]. I have my pacer in there to track the number of steps I walk. I have things on my calendar, what days I do certain exercises and stuff on there.” Another older patient, D11, a white man with a high school education, described a very different reason that his cell phone helped him to take control of his health, “They [providers] got our home phone number and there's times they call and leave a message on there, I've always got my cell phone with me and then if we're out for the day or something, I won't miss a message.” This patient used a cell phone to check his voicemail messages on his home phone while he was away. Although this patient may not be considered a typical example of a “high tech” self-manager, from the perspective of this patient, cell phones were explicitly prioritized for managing his health.
Many of these “proactive techies” placed a high value on health information, and especially prioritized scientific evidence to enhance accountability for acting on doctor recommendations. A20, a white woman with more than a 4-year degree, described the importance of health information and the value of email for enabling her to proactively communicate with her doctor about the accuracy of her research, "like "I read this about my child's condition, is this even accurate, would you agree, disagree?"...I need someone to double-check the information that I'm finding and give me an answer."

A12, a Hispanic woman with a high school education, also emphasized the value of health information. She expressed not having enough information, and wanted to learn more detailed scientific health information from her doctors, however: “They're [providers are] not prepared to have details about what are the studies behind this, what were the outcomes of that study and the percentages of side effects and that kind of thing. They're not ready at hand with that kind of thing.” This patient had the desire to have more scientific evidence for the health tests, medications, and behaviors that were recommended by her doctor. She recalled how, in the past, she had used different resources to get the information that her doctors could not provide, “The oncologist that I had connected with, I guess, told me what the options were. I researched what the options were and people who've gone through those options, and then came back with a decision on how to move forward with it.”

D6, a black woman with a 4-year degree, also described wanting more information so that she could be a better judge of her health status relative to the norms for her age and race,

“Unfortunately, they [providers] can't give you information you need on genetics and heredity and stuff like that, especially for African Americans, because we may be at some point of African descent but that's not all we are. For the majority of us, that's not all we are, so we're just grouped, and because this country has a 1/8 rule, we're just grouped... if you're African American and Asian your [A1C] numbers are higher because of the genetics there. That's what I'm getting at - so it really matters.”

Although six patients in this cluster were over the age of 60, and five had a high school education or less, all patients in this cluster used technology to be proactive about their health. Regardless of age, race, and ethnic background, these patients prioritized technology for tracking health indicators, communicating with providers, and finding scientific health information.

Cluster 2: “Indie Self-Managers”

Thirteen participants (5 male) define this cluster, including seven participants from the diabetes cohort (ages 62-89), and six from the asthma cohort (ages 28-43) (see Table 1). This cluster has high accountability for health self-management, expressing a strong belief that they have an active role in care (+3). Patients in this cluster considered technology to be a low priority for health. They are strongly independent (“indie”) and rely on keeping track of health tasks in their heads (+2). They disagree with the idea of needing follow-up from the doctor to stay on track with their health (-2), and with the idea that someone should check in with them to help remind them (-2). They strongly disagree that use of mobile technology is linked to better health outcomes (-3).

In contrast to cluster 1, patients in cluster 2 did not prioritize health information, nor did they connect technology use with health outcomes. Although all patients in this cluster described using a cell phone, and used it for other aspects of their lives, they did not consider it essential to their health.

For example, A4, a black woman with a high school education, said, “I use my cell phone and email and stuff, but I personally don't see how that affects my health.” Similarly, D8, an Asian man and the oldest patient in our sample, described why his cell phone was not linked to his health outcomes, “a cell phone is just another convenience, you know. I don't need it on the list of things I feel essential for managing my health. “A14, a white woman with more than a 4-year degree, made an even more explicit distinction between the use of her cell phones for everyday activities versus health activities. She described how her cell phone was directly at odds with the nature of her health self-management strategies, “my health needs me to be peaceful and relaxed and meditate and dealing with my emotional issues ... a cell phone's just a tiny little thing and it's actually a little bit detrimental. It's starting to become a little bit of a problem for me.” This patient noted the potential negative impacts on her health of communication technologies.
Although there was a weak link between technology and self-management for patients in this cluster, there was a strong sense of accountability. These patients felt strongly that they were responsible for their health, a statement that distinguished this cluster from others. D3, a black woman with more than a 4-year degree, gave a powerful example of how she took control of her health outcomes during a disagreement with her provider, “She [the doctor] wasn’t as receptive as I wanted her to be so I said some things at the end – it wasn’t in anger, but like ‘I’ll tell you,’ and because I said that she told me some things that’s been very beneficial to me.” Similarly, A14 described herself as an empowered patient, “I am confident, I have ownership. I feel like it’s my job to take care of that area of my life, I’m not passive, I’m an active player.” A9, a black woman with more than a 4-year degree, shared a similar sentiment about her role in care, “Just that, because it’s my body and I think I’m the expert. I know what’s going on with it versus someone that’s telling me what’s going on with it.”

Patients in this cluster prioritized health education because it helped them to be accountable for their health. The explicit connection between health education and accountability was unique for these patients. For example, A4 explained:

It empowers you, it holds you accountable versus thinking - I don't know, I've seen people who are like this is what's wrong, and they're constantly running to the doctor type of thing, but not really fixing what the issue is, just treating it constantly. I think it is good to be educated so that you could take care of it.

A9 had a similar view, “Getting the quality of the office visits we’re getting, the education that the doctor’s able to provide...It's nice to know what you're talking about and be able to help a child, based on what you actually know rather than guessing.”

For these independent self-managers, their “mental attitude,” as D8 put it, was more essential to health self-management than the technologies they used in everyday life. Their low adoption of technology specifically for health self-care was due to perceived negative impacts of reliance on cell phones, and an emphasis on autonomy from both technology and provider authority. They emphasized accountability and empowerment through health education.

**Cluster 3: “Remind Me! Non-Techies”**

Nine participants (2 men) define this cluster, including six from the diabetes cohort (ages 63-88) and three from the asthma cohort (ages 38-45) (see Table 1). This cluster has low accountability for self-management, and a low technology profile. They strongly agree that their health will improve if they set health goals with their doctor (+3), and that they need follow up from the doctor to stay on track with their health (+2). They dislike using electronic communication with their doctor (-1), and prefer phone calls (+1). They strongly disagree with keeping track of health tasks in their heads (-3), and do not believe in using online calendars (-1) or that mobile technology will improve health outcomes (-1). They agree that their health will improve if follow-up notices are clearer (+1).

Cluster 3 was characterized by their strongly positive attitudes toward goal-setting with their doctors and follow-up care. They placed a low priority on email and mobile technology and instead preferred phone calls with providers to stay on track with their health. Although they expressed agreement with being responsible for health, these patients were distinguished by their greater reliance on health care providers to guide and motivate healthy lifestyles.

The follow-up phone calls from providers were highly prized, if not relied upon, by these patients. For example, when asked about why she preferred phone communication, A18, a woman with some college education, said, “That's when I know I need to take action. Like if I've procrastinated for a while.” Similarly, A17, a white woman with more than a 4-year degree, found it difficult to prioritize health goals without the doctor checking in:

I get home and in my own brain, I sort of prioritize what was important and if life gets busy, I will prioritize other things above my health. If someone is calling you and holding you accountable to it, you feel if there's a priority for that person to call me, I'd better respect that, and keep it a priority.

Prioritizing health goals was a common challenge among these participants. A18 explained how follow-up could help her to stay on track, “If she [doctor] could have a goal, eating better or eating more fruits or whatever it is. And then once in a while if she sends me a thing that says ‘how's it going with this?’ – Oh yeah, I'm supposed to be doing this!”
Patients also perceived follow-up as an important source of motivation for staying on track with health goals. For example, D16, a black man with a high school education, talked about needing motivation from his provider. When asked how his doctor could motivate him, he said “Be there on me...My doctor's cell phone, he can do more.” This patient had several health complications, including two leg amputations, and limited speech due to a stroke. D5, a Hispanic woman with a high school education, who also had complications from comorbidities, described how better follow-up from her doctor could motivate her health management, “He'd [the doctor] make sure I go to the gym. He'd make sure I'm getting dialysis. He'd make sure he checks me once a month, they check everything I have to be doing.”

Finally, D14, a white woman with a high school education, expressed the value of having health goals with a provider, “If you go ‘that's the way life is,’ that's not helping anybody, but if he [the doctor] sets a goal – ‘okay, do more exercise, drink more water,’ then you know - I've got goals that will help me.” This echoes a statement from D2, a black woman with more than a 4-year degree, about the importance of health goals, “I think everybody needs goals, whether health or in anything, and goals would be a motivating factor.”

Discussion
The most surprising finding from this study is that it revealed attitude clusters that each contained a broad spectrum of patients in terms of race, age, education, patient-portal use, and diagnosis. Studies confirm that providers rely on these kinds of observable cues, especially race and age, to assess patients’ health attitudes and make treatment recommendations. In contrast, our findings provide evidence that patients’ attitudes toward health information and technology are not correlated with any of those observable cues. These attitudes and values are more influential than demographic characteristics on health self-management behaviors.

Our work can inform new approaches to clustering and understanding patients to inform tailored self-management interventions. In particular, our work suggests that attitudes toward health and technology can interact in unexpected ways, such as the “Indie Self-Mangers” who engage with technologies in everyday life but do not connect their technology competency with health self-management. Another example of unexpected interactions of health and technology attitudes is in “Cluster 1: Proactive Techies”, wherein six patients over the age of 60 years agreed that electronic communication and cell phones were essential to managing their health. Furthermore, A12 and A19, both women of color with high school educations, were “Proactive Techies” who expressed a strong sense of self-efficacy and technology use for health self-management. These examples are especially significant given that socio-economic status, which encompasses education status, has been shown to have the broadest effect on physician’s perceptions of patient competence, compared to race and age.

Researchers have extensively explored effects of race, gender, and literacy on clinician behavior, but have not investigated differences in the attitudes of a diverse set of patients’ toward self-management and technology. We found that patients’ attitudes have a strong influence on technology adoption and accountability for health that have implications for patient-provider relationships. In addition, those relationships extend through at-home, patient-centered self-management programs. For example, the “Remind Me! Non-Techies” had a passive attitude toward health self-care, and relied upon providers for reminders and monitoring of progress toward health goals. Moreover, members of this cluster contradicted stereotypes of white patients with high education, who are often perceived by providers as highly motivated and effective self-managers. For example, A17, a white woman with more than a four-year degree, did not use the patient portal, and relied on the doctor to keep her on track, “if the doctor's office is calling, you feel like oh! – okay.” Our results suggest that self-management programs for these low-technology users could leverage their respect for provider authority by delivering self-management content that encourages the feeling of being mentored by a provider. In contrast, interventions for “Proactive Techies” are more likely to be adopted if they support patient-initiated exploratory approaches to health tracking and goal-setting. Similarly, interventions for “Indie Self-Mangers” can be tailored to suit their strong preference for control over decision-making in care, providing tools that encourage autonomy and self-improvement through health education.

Finally, our work suggests that patients’ attitudes toward health information and technology may be important for understanding why patients adhere to, or fail to adhere to, treatment recommendations. Differences in these attitudes may have implications for health outcomes because these attitudes affect patients’ participation in care. Moreover, understanding the attitudes that distinguish different approaches to self-management can help providers to develop nuanced approaches to tailoring interventions for patients across a spectrum, from fully "activated" to fully "passive." Traditional scales that provide insight into patients’ perceptions of their role in care, like the PAM, can
be complemented with methods like ours that provide insight into patients’ strategies and motivations for developing their self-care capacity. We demonstrate the advantages of observing the interaction of health and technology attitudes for providing a holistic view of the attitudes and tool-use strategies that influence self-management.

Conclusion
Our findings demonstrate three unique clusters of self-managers based on different health and technology values. These three clusters demonstrate the limitations of a “one-size fits all” approach to self-management interventions, especially when these interventions are mediated by electronic communication tools like cell phones and email. We demonstrate that Q-methodology can be a useful approach for understanding patients’ attitudes toward technology and self-management. In addition, it could inform new approaches to clustering patients and matching them with appropriate self-care support. Prior research shows that raising clinician awareness of biases mitigates the effect of those biases, thus reducing healthcare disparities\textsuperscript{16}. This study suggests that an understanding of patients’ attitudes toward self-management and technology could inform unbiased, patient-centered interventions. Moreover, our approach has highlighted different values that influence technology adoption for health self-management, and this work could inform new scale development for assessing the appropriateness of computer-mediated interventions for patients. Overall, a greater awareness of the health and technology priorities of patients can help clinicians to leverage patient’s values and strategies for enhancing self-management. This work encourages clinicians and technology developers to combat implicit biases based on race, age, and education by eliciting patients’ values that define health self-management.

Acknowledgements
This project was funded by grant #R01HS021590 from the Agency for Healthcare Research and Quality (AHRQ).

References
41. Johnson RB. Examining the validity structure of qualitative research. Education. 1997;118(2).
Understanding the acceptance factors of an Hospital Information System: evidence from a French University Hospital

R. Ologeanu-Taddei, PhD¹, D. Morquin, MD¹², H. Domingo¹², MSM, R. Bourret, PhD²
¹Montpellier Research Management (MRM), Montpellier University, France; ²Montpellier University Hospital, Montpellier, France.

Abstract
The goal of this study was to examine the perceived usefulness, the perceived ease of use and the perceived behavioral control of a Hospital Information System (HIS) for the care staff. We administrated a questionnaire composed of open-end and closed questions, based on the main concepts of Technology Acceptance Model. As results, the perceived usefulness, ease of use and behavioral control (self-efficacy and organizational support) are correlated with medical occupations. As an example, we found that a half of the medical secretaries consider the HIS is ease of use, at the opposite to the anesthesiologists, surgeons and physicians. Medical secretaries reported also the highest rate of PBC and a high rate of PU. Pharmacists reported the highest rate of PU but a low rate of PBC, which is similar to the rate of the surgeons and physicians. Content analysis of open questions highlights factors influencing these constructs: ergonomics, errors in the documenting process, insufficient compatibility with the medical department or the occupational group. Consequently, we suggest that the gap between the perceptions of the different occupational groups may be explained by the use of different modules and by interdependency of the care stare staff.

Introduction
The implementation of Healthcare Information Systems, such as Computerized Physicians Order Entry (CPOE), Clinical Information Systems (CIS) or Electronic Medical Records (EMR) and, more recently, Hospital Information System (HIS), is supposed to have various benefits for the medical practices, as providing easy access to documentation of patients records and accurate them¹², billing management³, reducing potential medical errors⁴, and improving the quality of patient care⁵. However, previous studies have shown the use of HIS has led to unintended consequences in the actual work practices, such as increased documentation time⁶,⁷, incompatibility with clinical workflow⁸, increasing more interruptions in medical work⁹ and system-introduced errors in patients care¹⁰. How can we explain the gap between the expected benefits and the conclusions of these different studies? We suggest an explanation related to the fact that the acceptance factors are not taken enough into account by the hospital management during the post implementation stage. We argue this idea based on a survey driven in a French University hospital using a HIS. This survey is based on the main concepts of the Technology Acceptance Model (TAM)⁹.

Theoretical framework
TAM is considered a “gold standard” in the information systems research¹⁰. This model aims to identify determinants (perceived ease of use and usefulness) of attitude and usage intention. While it provides useful insights into why people adopt a technology, this model and its evolutions (known as TAM2 and TAM3¹¹), has several limitations. First, it works especially in voluntary contexts and not mandatory ones¹²,¹³. Thus, intention to use is not a relevant concept in mandatory contexts.
Second, it assumes that users face no impediments in the course of information system uses. Besides, users often do not have a total control over situations of use, especially as employees. This argument led TAM authors to improvement in the initial model¹¹.
Third, the parsimony of the TAM concepts, whereas it was considered for a long time as a strengthens of the model, has been evaluated recently as weakness because its misfit to different organizational contexts¹⁴.
Fourth, these concepts are linked to binary variables (use or not use), which may be described for simple technologies uses⁹,¹⁵ but not for complex ones. However, complex information systems, including Enterprise Resource Planning (ERP) and, consequently, HIS, give rise to more or less advanced uses. Some features may be used, while others not. In addition, the professional context of usage is not considered enough¹⁶.
Fifth, the main potential benefit of TAM, which is its explanatory and predictive power, seems to be partial. Sixth, previous research in the healthcare, based on the TAM use and focused on physicians, have found a positive relationship between perceived utility and acceptance but a non-significant relationship between ease of use and acceptance or ease of use and intention to use.

We suggest that these limitations are related to the deterministic use of the TAM, that is, measures of causality relations between the determinants, the attitude of acceptance and the intention of use. Beyond this deterministic use, the main construct of TAM may provide better understanding about HIS acceptance. In order to take into account the criticism we have summarized above, we add the construct of perceived behavioral control, which is driven by the theory of the planned behavior and has been used in the healthcare. This construct includes perceived internal control, which is named self-efficacy by the TAM authors, and perceived external control, which is linked to another concept of TAM, that is, facilitating conditions.

The table below summarizes the constructs of the survey and their definitions.

<table>
<thead>
<tr>
<th>Perceived Utility (PEU)</th>
<th>“the degree to which a person believes that using a particular system would enhance his or her job performance”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Ease of Use (PEOU)</td>
<td>“the degree to which a person believes that using a particular system would be free of effort”</td>
</tr>
<tr>
<td>Perceived Behavioral Control (PBC)</td>
<td>“The degree to which an individual believes that he or she has the ability to perform specific task/job using computer” (self-efficacy)</td>
</tr>
<tr>
<td></td>
<td>“The degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system.” (facilitating conditions)</td>
</tr>
</tbody>
</table>

Table 1. Definitions of the constructs

Methodology and context

We designed a questionnaire based on these three constructs. We used the items tested by Chau & Hu. All items were measured in a 7-point Likert’s scale, with 1 as strongly disagree and 7 as strongly agree. We added two open questions in order to investigate users’ perceptions about the advantages and the disadvantages of this information system.

This questionnaire was sent for pre-test to a panel of 20 volunteers. This pre-test and adaptation of the questionnaire prevented the definition of indicators according to the representations of researchers and thus allowed the contextualization of the issues. The main adaptation consists on the contextualization of the HIS name: as suggested by the pre-test answers, we decided to use the software’s name instead of the generic term “HIS”.

The questionnaire was developed and administered online to the care staff, during the month of December 2013. The survey was conducted in a large French University hospital. The target was composed by the care staff (9 000 employees with care occupations). The aim was to measure the users’ perceptions related to the daily use of the HIS, 18 months after the HIS implementation. This information system includes computerized physicians order entry, medical and nursing observation, laboratory tests results, medical prescription, operating room process management, drugs logistics management, consultation program and appointment, billing management.
Data analysis and results

1,942 questionnaires were collected. The distribution of the occupations is detailed in the table below:

<table>
<thead>
<tr>
<th>Occupation</th>
<th>No. Quot.</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurses</td>
<td>656</td>
<td>33.8%</td>
</tr>
<tr>
<td>Physicians</td>
<td>327</td>
<td>16.8%</td>
</tr>
<tr>
<td>Medical secretaries</td>
<td>169</td>
<td>8.7%</td>
</tr>
<tr>
<td>Others</td>
<td>163</td>
<td>8.4%</td>
</tr>
<tr>
<td>Residents</td>
<td>117</td>
<td>6.0%</td>
</tr>
<tr>
<td>Auxiliary nurses</td>
<td>103</td>
<td>5.3%</td>
</tr>
<tr>
<td>Nurses managers</td>
<td>79</td>
<td>4.1%</td>
</tr>
<tr>
<td>Anesthesiologists</td>
<td>73</td>
<td>3.8%</td>
</tr>
<tr>
<td>Surgeons</td>
<td>56</td>
<td>2.9%</td>
</tr>
<tr>
<td>Physiotherapists</td>
<td>38</td>
<td>2.0%</td>
</tr>
<tr>
<td>Pharmacists</td>
<td>35</td>
<td>1.8%</td>
</tr>
<tr>
<td>Social workers</td>
<td>32</td>
<td>1.7%</td>
</tr>
<tr>
<td>Midwives</td>
<td>26</td>
<td>1.3%</td>
</tr>
<tr>
<td>Psychologists</td>
<td>24</td>
<td>1.2%</td>
</tr>
<tr>
<td>Pharmacist residents</td>
<td>20</td>
<td>1.0%</td>
</tr>
<tr>
<td>No answer</td>
<td>15</td>
<td>0.8%</td>
</tr>
<tr>
<td>Dentists</td>
<td>9</td>
<td>0.5%</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>1942</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 1: Occupational characteristics

We use the across-method triangulation, consisting in the combination of three methods:

- Statistical correlations between the responses to the closed questions, which were assessed by Chi-Square test, using the Sphinx software.
- Content analysis using a French textual data analysis software named Alceste, which identifies classes of speech.
- Content analysis of the open answers, which were coded independently by two researchers (the reliability was measured by the kappa coefficient).

We also used internal validation, which consists in the “truth value” of the findings and their credibility for the care staff. We presented the findings to the hospital management, to the medical managers of the hospital units or services and to the care staff.

We found a very significant correlation of all the constructs with occupations. We resumes bellow the main rates (1-3 for disagree and 5-7 for agree. We do not mention the middle rate (4 on the Lickert’s scaledivided):

1. **Item Perceived Ease of Use (PEOU):** \( \chi^2 = 160.48, \quad ddf = 30, \quad P < 0.01, \quad Cramer’s V : 20.42\% \). Generally, 48% of the staff care agree (vs. 35% disagree) with the statements related to the construct “the HIS is ease of use”. The majority of anesthesiologists (61%), of physicians (56%), of midwives (55%) and of surgeons (52%), and of the residents (51%), the half of nurses, 45% of care managers, 39% of auxiliary nurses, 35% of physiotherapists disagree. Only 25% of anesthesiologists, 27% of physicians, 27% of midwives, 30% of surgeons, 33% of nurses, 35% of residents, 37% of care managers, 42% of auxiliary nurses, 45% of physiotherapists agree with this statement. Furthermore, 50% of the medical secretaries (vs 32%) consider the HIS is ease of use. Pharmacists are more divided: 41% disagree and 40% agree.

2. **Item Perceived Utility (PU):** \( \chi^2 = 168.61, \quad ddf = 30, \quad P < 0.01, \quad Cramer’s V : 19.65\% \). The staff care consider the software is useful or very useful (42% agree vs. 34% disagree). More precisely, 48% of auxiliary nurses (vs. 31%), 38% of anesthesiologists (vs. 38%), 57% of care managers (vs.13%), 47% of surgeons (vs. 27%), 44% of residents (vs. 27%), 41% of physicians (vs. 38%), 43% of physiotherapists (vs. 27%), 74% of pharmacists (vs. 15%), 31% of midwives (vs. 23%), 63% of medical secretaries (vs. 16%)
agree with the statement “the HIS is useful”. Besides, nurses (30%), midwives (31%) and anesthesiologists (38%) have the lowest rate of PU. Only the nurses disagree (46% vs 31%) with this statement. In contrast, pharmacists (71%), medical secretaries (63%), residents (61%) and nurses managers (57%) state the software is useful or very useful.

3. **Item Perceived Behavioral Control (PBC):** chi² = 154.32, ddl = 30, 1-p = >99.99%. The staff care is divided (46% agree while 36% disagree), as surgeons (45% agree vs. 41% disagree). More precisely, 52% of anesthesiologists (vs. 32%) and 43% physicians (vs. 37%), 44% of pharmacists (vs. 39%), consider they have only little behavioral control or not at all. At the opposite, 65% of medical secretaries (vs. 20%), 64% of residents (vs. 34%), 63% of physiotherapists (vs. 22%), 56% of auxiliary nurses (vs. 25%), 45% of nurses (vs. 31%) and 42% of nurses managers (vs. 35%) states that they have behavioral control.

Furthermore, we identified 4 classes of speech by the Alceste software:

- **Class 1 (majority) "Treatment plan".** Key words: prescription - care - treatment sheet - validation - error - transmission - treatment - sign - readability. At the center of this class are prescription and treatment plan, related to the difficulty to understand the computerized prescription and treatment sheet, which may lead to errors of treatment.

- **Class 2 "Lack of ergonomics".** Key words: unfriendly - tool - person - think - Computer - true - Software - design. These speeches are related to the ergonomics software, considered as flawed. Some users describe "a disaster" loss of staff time to the detriment of his care tasks. Software bugs are also highlighted.

- **Class 3 "Laboratory results and examinations".** Key words: report - post - results - classify - archive - imaging. The speech emphasized results and schedule modules, which are described as unusable, as well as the non-readability of biological results.

- **Class 4: "Find and view information".** Key words: Input - Support - File - information - structure - Psychiatry - enter - check. This class focuses on the difficulties of entering and viewing documents external to the University Hospital and the difficulty to find relevant information. Many people feel that the information is misfiled.

The difficulties related to the HIS may be classified into three categories, as follows:

- **Software ergonomics.** The HIS is viewed as a usability-faulty system. Overload of information on the screen drowns relevant information and increase the risk of error. Users mention also bugs and software failures.

- **Insufficient use of documentation rules in the clinical workflow and misfiled information driving to errors in patient care.**

- **Insufficient compatibility with the clinical workflow in the departments (e.g. Hematology, Emergency, Pediatrics and especially Psychiatry) and for different medical occupations (anesthesiologists, physicians, surgeons, midwives), because the possibility of settings and customization was not effectively made.**

In addition, the words occurrences analysis shows that the mains words are: time – especially “waste of time” (727 occurrences); difficulties or difficult (402 occurrences); waste (383 occurrences); prescriptions (257 occurrences); incompatibility – of the information system to the clinical work (199 occurrences); prescription errors (182 occurrences); training (insufficient) (164 occurrences); errors – especially, bugs (154 occurrences).

**Discussion**

The low score of PEOU and PBC or the anesthesiologists, the surgeons and the physicians may be explained by the answers to the open questions. Based on the classes of speech, we have to consider the prescription module and the treatment plan module as the critical issues of the HIS use. The lack of ergonomics of the prescription and laboratory results modules, related to the information overload, may lead to prescription or treatment errors (especially for the drugs dosage and their frequency). In this context, anesthesiologists assess the lowest score of PEOU because they are the clinical occupation for which the medical error (especially prescription error) may have the most rapid dramatic consequences. Basically, anesthesiologists, physicians, surgeons and midwives are the most concerned by this issue, contrary to the medical secretaries.
Otherwise, the findings highlight errors in the documenting process associated with the misuse of indexation rules for patients’ files and the insufficient compatibility of the HIS to the clinical process for each medical department.

We have to mention that ergonomics, errors in the documenting process and the insufficient compatibility with specific medical processes have already been identified in the literature as a medical error risk. Our mixed method shows these factors are very connected. Furthermore, ergonomics and usability, correlated to a lower rate of PEOU, are a main issue, in contrast with studies suggesting physicians’ ability to learn to use the technology by themselves and their relative disinterest in usability. Moreover, the PBC rates suggest that the medical staff (especially anesthesiologists) considers that organizational support (as training) is not sufficient during the post-implementation stage. This result is consistent with the literature on the ERP success which is conditioned by the managerial and technical support.

In addition to these issues supported by the literature, our study adds to the literature the gap between the HIS perceptions according to the occupational groups. As an example, we found that a half of the medical secretaries consider the HIS is ease of use, at the opposite to the anesthesiologists, surgeons and physicians. Moreover, they reported the highest rate of PBC and a high rate of PU. It would be easy to conclude that this result is the expression a conflict between the administration and the professional logics, knowing that accountability and planning are the main aims associated to an ERP and, consequently, to a HIS. Nevertheless, an in-depth look at the results emphasizes a gap inside the clinic and care occupational groups perceptions. Thus, pharmacists reported the highest rate of PU but they are more skeptical about the PEOU and PBC. Answers to open questions suggest that this result is related to the drugs traceability, which is on expected benefit of the HIS, under the condition of the improvement of both the ergonomics and the indexation rules for patients’ files. Otherwise, the different clinical and care occupational groups do not need to use the same features. Anesthesiologists, surgeons and physicians use the prescription module, while nurses and auxiliary nurses use the care sheet. Nurses are very impacted by the misuses of the prescription module by the clinical staff, which may explain their low rate of PU and PEOU. In line with this argument, we suggest that the most nurses state that they have PBC because the right use of prescription module is not their task. Thus, we may consider our results in relation to the use of the different modules and features according to the occupational groups (e.g. use of the office module for medical secretaries, use of the prescription and the results laboratory modules for the physicians and surgeons) and the exposure to the medical liability and medical malpractice.

Although each occupational group uses different modules or feature of the HIS, they are interdependent, meaning that an error in the information workflow may have consequences on the clinical and care professionals involved in the workflow (related to a patient). Thus, medical secretaries have to use the indexation rules for the patients’ files for physicians, surgeons and anesthesiologists to find the relevant information related to the identity and the admission file. Clinical occupational groups have to use the right features for medical observation and prescription for pharmacists to deliver drugs and for nurses to administrate drugs and care delivery. Consequently, the PU, PEOU and PBC may influence these constructs for the other occupational groups.

**Conclusion and future directions**

Our study shows that perceived usefulness, ease of use and behavioral control are correlated to medical occupations. Furthermore, it outlines three main points that may explain this correlation: ergonomics, errors in the documenting process and insufficient compatibility with specific clinical care process. Generally, information systems are implemented according to a standard vision of clinical workflow. Our research results suggest that the HIS have to be settled and customized taking into account particular clinical care process. In addition, these results can help planners, managers and healthcare software editors to understand key issues affecting HIS perceptions and use.

This research provide better understanding inside the HIS care staff perceptions. Besides, the generic items of the questionnaire, driven from TAM, may be completed by more specific items considering each feature and module of the HIS (medical observations, prescriptions, care treatments, laboratory results), as suggested by the answers to the open questions. Indeed, medical secretaries, nurses managers, nurses or physicians do not use the same features; thus, theirs perceptions and their expectancy about the HIS are different.

Further research have to provide also more insight into the various features and modules used by the medical and the care employees, in different departments, and on their interdependency in both the information and care chain.
Acknowledgements

This study was supported by the Research Federal Structure of the Montpellier University Hospital and the Montpellier University.

References

ABSTRACT:
User frustration research has been one way of looking into clinicians’ experience with health information technology use and interaction. In order to understand how clinician frustration with Health Information Technology (HIT) use occurs, there is the need to explore Human-Computer Interaction (HCI) literature that addresses both frustration and HIT use. In the past three decades, HCI frustration research has increased and expanded. Researchers have done a lot of work to understand emotions, end-user frustration and affect. This paper uses a historical literature review approach to review the origins of emotion and frustration research and explore the research question; Does HCI research on frustration provide insights on clinicians’ frustration with HIT interfaces? From the literature review HCI research on emotion and frustration provides additional insights that can indeed help explain user frustration in HIT. Different approaches and HCI perspectives also help frame HIT user frustration research as well as inform HIT system design. The paper concludes with a suggested directions on how future design and research may take.

Author Keywords:
Frustration, Emotional design, Affective Computing, Health Information Technology (HIT), Human-Computer Interaction (HCI), Interface design, Clinicians, User Experience (UX).

ACM Classification Keywords:
H5.2. Information interfaces and presentation: User Interfaces.

INTRODUCTION:
One of the overarching goals of Human-Computer Interaction (HCI) research has been to increase the good experiences and possibly reduce or mitigate horrible first experiences with technology. Licklider, in his “Man-Computer Symbiosis” [14] article decades ago, envisioned how a perfect, and relatively mutual relationship between humans and computer systems should be. When we think of times or experiences when we had to use technology for the first time or switch to a different technology, the thought may not always be pleasant. In fact, mostly unpleasant. Generally speaking, depending on one’s generation and “mental model”
of the new technology, the thought of this experience will generate either a happy or not so happy feeling. When technology is used, regardless of the type of technology used and reason for use, there is a formation of an initial relationship with the technology. Should the initial relationship last, a long-term relationship is sustained as well. Frustration which has long been defined by Lawson [25] as an “emotional state resulting from the occurrence of an obstacle that prevents the satisfaction of a need”, summarizes what the lasting relationship may entail.

RESEARCH PROBLEM:

As the American healthcare system slowly moves through technology adoption, implementation and meaningful use of Health Information Technology (HIT), one of the main areas of concern is the Human-Computer Interaction (HCI) and user experience of clinicians as they adopt, implement and use HIT. The problem here is that there has been research showing that frustration is one of the top unintended consequences that occur during Electronic Health Records (EHR) implementation and also through the first few months of post-implementation usage [5, 21, 22, and 25]. With literature repeatedly confirming the incidence of frustration in HIT system usage, there should be work done to finding successful approaches to solving frustration in HIT interfaces. This paper attempts to use such HCI research to explain and possibly inform design that could reduce frustration in HIT use.

ORIGINS AND EARLY RESEARCH ON END-USER FRUSTRATION:

With this concept of human-technology relationships, HCI research has re-defined itself to explore first experience interactions to find interdependencies, shared workload and address different emotional states such as frustration in technology use. HCI research has gone into issues such as why humans develop emotions towards technology and how to mitigate negative emotions towards technology. Domains like emotional design, affective computing and frustration research are born out of researching into these kinds of problems.

One of the main HCI research disciplines for some time now has been exploring ways in which human feelings towards technology can defined, quantified and monitored. The basic goal of this field is to know more about HCI and how to sustain and maintain this ‘relationship’ as ethically as possible [1]. The beginning of a shift in focus of emotion design and end-user frustration research started with cognitive and social research. Before the 2000s, research was already being done on frustration, affective computing as well as studying emotional states. Researchers such as Rosalind Picard and the likes, have spent years in Massachusetts Institute of technology (MIT) Media Lab studying the recognition, measurement and description of affect [9, 11, and 12]. Her work over time, especially on affective computing explores how computing relates to emotion.

RESEARCH IN THE 2000s:
The early 2000s saw research on emotional design such as works by Don Norman [17, 18, and 19]. Don Norman’s article “Emotional design: Why we love (or hate) everyday things” and his book on “Emotional Design” started a way of thinking that suggested that good design was imperative and that bad design not only generates a negative feeling like frustration but is also was not cost effective. Norman also collaborated with Ortony in “Designers and users: Two perspectives on emotion and design”. The goal of this article was to find out and share perspectives of emotion and design by both designers of technology and users of technology [19]. This research explains how a shared needs analysis between designers and users is essential to understanding what emotions are and how emotion is experienced with technology.

Lazar et al[13], also did some work that explored frustration focusing on what influenced a frustration experience. They wanted to find out more about causes and effects of frustration and so they looked at a bunch of users’ frustration events and reported on incidental (happening at one time) events versus session (over a period) events. These frustration events were the generic HCI frustrations or annoyances experienced when using a computer anywhere and were measured in levels. In the research, the authors found out that although a lot of user factors accounted for frustration, more people are angry at the computer than they were angry at themselves or even determined to fix the problem. Whether or not a frustration event was incidental or session was really critical in terms of the rating of frustration severity. Findings also included the fact that the user’s mood before and after a frustration event played a huge role in frustration severity.

In terms of emotional design research, round about the same time as Lazar and team were working on frustration experience research, Boehner et al [4] put out a paper that was unlike the then current approaches which defined emotion as just a cognitive experience. They argued that emotion was not just a state of being but interactional as well. Emotion information could be captured processed and output just like Card et al’s Human Processor Model. This argument and their research as a whole not only made it easier to break down emotion information and measure emotion in the 3 phases of the IPO model( i.e. during input, processing and output) but their explanation of how emotion can be treated as both information and interaction helps define, understand and explain emotion as well as measure it. Emotion can be socially and culturally understood and interpreted and this means that we should be able to understand and define emotion as both internal and also expressed differently in different contexts by different people.

Two years before this research, Boehner et al [3] did some research on affect. In their paper “Affect: From Information to Interaction”. Similar to their work on emotion as information, they explained how affect in affective computing approached as merely processed and expressed as information leaves no room for the feeling part of affect. Their approach (interactional) was rather from a perspective where the emotion or human users in affective computing is not lost but understood and supported as well. They came up with reasons to support why and how the interactional approach will work better. In essence when you approach affective computing and define affect as interaction as well, then one is able to see affect from both a social and cultural
product which relies on and supports flexibility and is not confined or formalized. Their work also acknowledged some challenges in this kind of approach however, the understanding that emotion is one way of looking into interactions helps think about ways to design for a more broader population and use. Bickmore and Picard’s research article “Establishing and maintaining long-term Human-Computer Relationships” in 2005 expanded on what human–computer relationships were in the first place.

Their goal was to find out how personal and interpersonal social relationship constructs can be applied to human-computer relationships. It is interesting to note that they were able to identify that the same reasons for relationships in human-human interaction could be applied to human-computer interaction. So for example, people could form relationships with ATM machines for economic gains just as people form relationships with other people for economic gains. Human relationship communication and management also applied to human-computer relationships. The authors also introduced relational agents to help with long term human-computer relationships that focused on and sustained despite multiple interactions. In research like this, there is a new window opened for HCI designers and researchers to look into. It is easier to understand human-computer relationships if we can compare it with human relationships. User frustration research was still being done.

Research on general computer usage and frustrating experiences like those by Ceaparu et al[7,8] measured how often frustration occurred, what caused frustration and how bad the experience was. These research looking at user experiences on common applications on web platforms and how users handled it. They recorded that some common experiences were during mailing and word processes. The perceived causes of end–user frustration ranged from error messages to dropped network connections to hard-to-find features on interface. These research showed that frustration is a huge problem that needed to be addressed. Based on the data they received, it was evident that self-reported high levels of frustration and wasted time was among the top list of concerns in a frustration occurrence. These data also could be expanded into both financial and usage ramifications in the future. Knowing these causes, and frequencies of user frustration, research is a critical first step in exploring more about what could be done to reduce user frustration. Other researchers also studied how computer systems and interfaces could automatically detect frustration in users [11] and also how to design computers to respond to user frustration [12].

**RECENT RESEARCH ON HCI FRUSTRATION:**

More recently, research on frustration has focused on more dynamic aspects on end-user frustration like how we can harness frustration for the good. In fact, Frustration, sometimes referred to as confusion or anxiety or anger, is a special type of negative feeling or emotional state experienced not only in life but also in HCI. This is described as “A strong, negative emotional state” that users feel when they interact with technology or computer system interfaces [7, 9]. Frustration in HCI has been considered to hinder system
usability. For example, the research by Lazar et al [13] showed that there is a difference in frustration experience in student users versus worker users of a computer system. Which means a more tailored approach can be researched into so that narrowly focused research and design suggestions can be made for EHR design so as to mitigate frustration.

**A little bit of frustration is good!**

Baker et al [1] in their article on frustration compare the “cognitive-affective” states of students in different computer-based learning environments. Amongst other things, they found out that frustration was better than boredom in the different kinds of “cognitive-affective” states they evaluated. It can be inferred from the article that in any learning environment like what clinicians experience during the implementation of EHRs, confusion and a little bit of frustration may be better for the learning experience than boredom. This meant that it will be prudent to invest in system designs that randomly interrupted with routine boring processes but also monitor the amount of interference so as not to create a terrible experience for clinicians.

Riseberg et al [24] had long before researched on using frustration for the good. In their article “Frustrating the User on Purpose: Using Bio-signals in a Pilot Study to Detect the User's Emotional State”, they show how frustration could be induced in computer game interfaces in order to be able to detect and group frustration events that create similar emotional states making it a great approach to explore users’ states during human computer interaction. Current research [9, 10] refine what Riseberg et al did. Mentis [16] also did some work on interruptions in cognitive workflow that elicits the experience of frustration. She found out that in instances where people remember their frustration moments in interface interactions, those moments were times where the system had to respond to user needs/actions and not times when users were figuring out how to translate their thoughts to actions on the interface. This means that remembered frustration experienced occurred in the second half of an interaction mostly because there was a break in what interfaces were supposed to be displaying or executing to users.

**WHY HCI LITERATURE REVIEW IS NEEDED:**
The HCI literature reviewed provides an understanding of some basic parts to frustration as well as emotional design research. The literature review explores origins of end-user frustration, what has been done quite recently and future directions that researchers could take. From the literature, we know that man and computers can form a symbiotic relationship in which each contributes to the relationship [14]. This means that it is essential to know which parts of an HCI that humans take responsibility versus which parts a computer should be left to complete. Not only do we know that such human-computer relationships can exist we also know that we can apply human-human concepts, definitions and constructs to human-computer relationships as well [2, 6]. The same way human relationships involve emotions and sometimes frustrations, we are able to understand that there is the need to design for emotions and cater to frustration events should they arise [2]. User-frustration research which can also be found in affective computing research championed by people like Picard, Norman, Dourish, Bohner, Schneiderman etc. have been able to narrow down research to reflect specific
interface use, human factors, system factors, and perspectives. This is very helpful because then in thinking about using HCI use-frustration to explain and possibly influence future research and design in HIT, I will consider such finding and use their results as baselines to further research in HIT user frustration.

From the literature, we know what the right questions should be. For example, we know that the question in user frustration is not whether frustration is a good emotional feeling or not. We even know that in certain contexts, a little bit of frustration is good [1] and frustration has been used to find out more about emotional states [9, 24]. However, in thinking about what contributes to frustration and what HCI concepts helps better understand frustrations in HIT interfaces, a lot of factors come to play; long-term HCI relationship[2], definition of the affective state called frustration (information or interaction)[3,4], severity of frustration(self-reported or actual)[7,8], implications of severity of frustration in a system(economic, social, usage, perceived)[8,12], perspectives(system, user or designer) [11, 12, 13, 19], ways to measure user frustration with systems[9,11,16,22] come into the picture.

The solution to HIT clinician frustration is to first identify what makes clinician frustration different from the frustration events described in the literature. Specifically, what is causing frustration at that moment? From Bickmore and Picard’s work [2], it is safe to assume that overarching goal of any HIT implemented is to maintain a long-term relationship between clinicians and the systems after implementation. With the HITECH act and meaningful use, there is even a higher incentive to mitigate frustration which hinders such long-term relationships. This a broader picture of why frustration hampers the goal of IT adoption in healthcare.

There is also the issue of context and expectations from the clinician. In trying to address frustration, it will be better to consider what role the HIT is playing and the consequences of interruptions such as frustration. This means that technology which most ideally satisfies its reason for being, is subtle. Just like breathing, we wouldn’t know we are using it. He further explains that this evolution of technology adoption and implementation is only natural “a fundamental consequence not of technology but of human psychology....Whenever people learn something sufficiently well, they cease to be aware of it.” There is therefore a presumption that the initial phases of computer or technology interactions are not as unconscious or spontaneous as later phases where the technology is pervasive. Our experiences with technology, especially for the first times are memorable; either in a good way or a horrible one.

**Combined-approach to frustration research in HIT:**

In any HCI, there is a human and a computer system involved. A combined approach to frustration research in HIT will be to combine the human-side research with the computer-side design approach to mitigate frustration. In this case, issues like shared interaction and designing for human users are discussed jointly with HCI themes and theories. Interaction theories may be in the form of visualization, peripheral vision, affordances, human cognition, user preferences, abilities, capabilities, limitations, mental models, etc.
DISCUSSIONS AND FUTURE RESEARCH:

Over the timeline of frustration in HCI research, there is much discovered. We know what frustration is, what causes it, how we can measure it, how we can attempt to mitigate it both from user side and system design side, we know what other human states (pscho-/physio-/socio-, cognitive, mental, effective, affective etc. etc.) are involved in HCI frustration. The next phase of research should probably focus also on prior experiences that feed onto frustration and how that better tells the story of how frustration is experienced.

There is also the need to critically research into what makes frustration in HIT hypothetically more severe or relatively unbearable that frustration in other interactions. We may (or may not) discover that the frustration experience is just what it is regardless of the kind of technology or interface interaction. Assuming HIT interaction frustration proves more annoying (due to many other factors like safety, lives at stake, timeline crunch etc.), then research much begin to address it in a more ‘wholistic’ way like the combined approach. Perhaps it will be best to approach it will be by framing a scenario (see fig 1).

**Fig1: HIT frustration incidence model.**

**Framing a Scenario:**

This scenario is just a way to approach an instance of frustration that can be comparable to a use case in system design. Assuming you have a nurse¹ experiencing frustration with a computer order entry (CPOE) interface, what could be a logical HCI research-based explanation to such experience? We know that frustration can happen in all stages of technology use especially the initial phases.

To frame this scenario and better explain the frustration; let’s break the experience into three (3) phases: Pre-experience, inter-experience, and Post-experiences;

- **Pre-experience Phase:** The pre-experience phase of any HCI is the collection of knowledge, experience, ideas, mental models and skills that a user already has before he or she experiences frustration. Most at times, stress which is a higher determinant of frustration may play a major role in inducing frustration

  ¹Note that this model may illustrate several frustration circumstances.
during interaction. In our scenario, the nurse’s background, prior experience(or not) with interface, values and beliefs, physical and intellectual capabilities and time constraints could factor into her pre-experience that leads to frustration.

- **Inter-experience Phase:** This phase will be the particular moment when the clinician is in front of the interface interacting with the computer. Here the clinician’s physical, social and cognitive abilities, capabilities and limitations influence the severity, instance and overall frustration experience at the moment. The interface’s design, visualization features, affordances, display and feedback also contribute to the experience.

- **Post-experience:** After the frustration experience, a clinician acts upon one of three feelings (see fig 1) that they are left with. 1. A self-motivating feeling that makes them want to try again and not give up. 2. An indifferent feeling that makes him/her only use the technology or interface when necessary but avoid the interaction most of the time. A perfect instance is when clinicians are required to use hospital technology. 3. A worst-case scenario when a clinician gives up and decides not to use the interface any more.

These 3 post-frustration decisions are critical in that whichever one a clinician decides on: a) Determines the frustration threshold they have reached. b) Determines what mitigation steps to use. c) Could lead to insights on where the link broke.

**CONCLUSIONS:**

When user frustration is explained using HCI principles and themes, there will be a better way to address the issue. The US healthcare system still faces some push-back in adoption and use of HIT. Clinicians are frustrated by EHR systems; either by interface design or how these systems interrupt workflow. The goal is to move past current research into what frustration is, to applied research on how the knowledge of frustration can be applied to HIT. Unlike frustration in other interfaces like web navigation, HIT interfaces has the component of frustration by limited interface view.

There are other distinct differences that can be explored. Frustration in EHR interfaces occurs at all stages especially the initial implementation phase. When a clinician experiences frustration; there are three (3) phases involved: pre-experience, inter-experience and post-experience. All these three (3) phases have factors (especially HCI theories) that come to play. Health care interfaces may have certain distinct features or functions that makes frustration experiences in such HCI’s different.

Our interactions with the technology should be able to amount to a relationship that can be sustained rather that a relationship that makes it difficult to maintain due to experiences such as frustration.
ACKNOWLEDGMENTS:

I would like to thank Helena Mentis PhD, mentoring me whiles I explore this topic and also for her advice on drafting this paper.

REFERENCES:


Identifying the Clinical Laboratory Tests from Unspecified “Other Lab Test” Data for Secondary Use

Xuequn Pan, PhD¹,², James J. Cimino, MD¹,²
¹Lister Hill National Center for Biomedical Communications, National Library of Medicine; ²Laboratory for Informatics Development, NIH Clinical Center; Bethesda, MD

Abstract

Clinical laboratory results are stored in electronic health records (EHRs) as structured data coded with local or standard terms. However, laboratory tests that are performed at outside laboratories are often simply labeled “outside test” or something similar, with the actual test name in a free-text result or comment field. After being aggregated into clinical data repositories, these ambiguous labels impede the retrieval of specific test results. We present a general multi-step solution that can facilitate the identification, standardization, reconciliation, and transformation of such test results. We applied our approach to data in the NIH Biomedical Translational Research Information System (BTRIS) to identify laboratory tests, map comment values to the LOINC codes that will be incorporated into our Research Entities Dictionary (RED), and develop a reference table that can be used in the EHR data extract-transform-load (ETL) process.

Introduction

Laboratory test results are stored in electronic health records (EHRs) as structured data coded with local or standard terms. These formally coded laboratory tests can facilitate retrieval and reuse of EHR data. However, laboratory tests conducted at an outside laboratory are often simply labeled or coded as “outside test” or something similar, with the actual name included in the free text result, comment, or note (e.g. “Test requested: Lyme Disease Serology Test performed at: Mayo Medical Laboratories Rochester, MN Test result: Lyme Disease Serology, S - Negative”). As a result, these outside laboratory tests with nonspecific names cannot be differentiated during retrieval, impeding tasks such as patient care, data sharing, integration, analysis, and decision support. Manual clarification of such data is tedious and redundant.

Our goal is to develop a generalized method to make outside unspecific laboratory data available for secondary use. Our approach seeks to code nonspecific tests with appropriate codes and standard terms from the Logical Observation Identifiers Names and Codes (LOINC) using a fuzzy matching approach that comprises four steps: 1) identify outside laboratory tests results, 2) based on the text fields, map the test to a specific standard LOINC code, 3) develop local codes, 4) recode outside unspecific laboratory test results proactively when loading new data into the EHR. We demonstrate this approach with the Biomedical Translational Research Information System (BTRIS), a repository of EHR data at the National Institutes of Health (NIH).

Background

Logical Observation Identifiers, Names and Codes (LOINC)

LOINC, a standard for reporting clinical observations (including laboratory test results) used in EHR systems, includes names and identifiers for more than 68,000 medical terms. Approximately 300 LOINC codes cover more than 95% of laboratory test orders in the U.S. Based upon available laboratory information (e.g. name, unit of measure), we can use the LOINC Mapping Assistant (RELMA) to match LOINC codes with local codes. Over the past two decades, the benefits of LOINC mapping for EHR interoperability have been well discussed, as have the challenges in mapping practice. Basic LOINC mapping guidelines are derived from an accumulated experience with MIMIC-II lab codes.
The Biomedical Translational Research Information System (BTRIS)

BTRIS is a clinical research data repository at the National Institutes of Health (NIH) that collects EHR data from over 50 NIH sources. Laboratory data are obtained from multiple systems including the current hospital laboratory system, archived data from previous systems, and institutional clinical trials management systems. All of these sources include results from outside laboratories (e.g. Mayo Medical Laboratories).

All data in BTRIS are coded with the NIH’s Research Entities Dictionary (RED), a terminology resource that includes the concepts related to laboratory tests or panels in five categories (chemistry, hematology, immunology/flow cytometry, microbiology, and transfusion medicine). Like other RED concepts, laboratory test concepts have properties, roles, and associations to represent comprehensive knowledge about the tests they represent.

BTRIS uses a hybrid relational and Entity–Attribute–Value (EAV) database model, in which most laboratory data are represented with columns in two tables: Event_Measurable and Observation_Measurable. The Event_Measurable table contains laboratory order information with orderable laboratory tests and panels stored in the Event_Name column along with their RED concepts in the Event_Name_CONCEPT column. Laboratory test results are recorded in the Observation_Measurable table. Laboratory finding values (results) are in Observation_Value_Text or Observation_Value_Numeric columns. Additional textual information is stored in Observation_Note and Observation_Value_Name columns. Figure 1 displays a sample of the two tables. Additional information (e.g. status flags and reporting time) are stored in the corresponding EVA tables.

Figure 1. Sample data on the Event_Measurable and Observation_Measurable tables (selected columns only).

Fuzzy Lookup Transformation

Fuzzy matching or lookup uses mathematical processes to determine the similarities between strings and to find similar (non-exact) matches. We expect this method to be sufficient for matching newly received results with similar previously encountered results, since they typically share similar patterns and formats. Fuzzy matching can be expected to overcome minor differences such as unique numeric results, misspellings, or inconsistent abbreviation usage. For example “Aspergillus fumigatus Ab IgE 78.4 ku/l laboratory” and “Aspergillus Fumigatus Aby Ige 36 Ku/L” will be readily recognized as matching.

The development of fuzzy technology has a wide range of potential uses in health care (e.g. diagnosis, decision making); its role in EHR data cleaning and standardization is relevant to the current project. EHR data extract-transform-load (ETL) processes consolidate disparate clinical data into a data repository. Microsoft SQL Server Integration Services (SSIS) features a fuzzy lookup transformation in the ETL package for data cleaning and standardization. We plan to use the SSIS platform to perform fuzzy matching to recode unspecific laboratory data and incorporate it into the ETL processes to check new data during real-time data loading.

Methods

As previously described, the nonspecific laboratory tests are those conducted at an outside laboratory and the results are stored in clinical databases without specific test names. Our goal is to develop a generalized method that makes outside unspecific laboratory data available for secondary use. The following illustrates our four-step approach.

Step 1: Identifying nonspecific outside laboratory tests

To find unspecified outside (or reference) laboratory tests, we looked in the RED for suspicious local terms referring to outside laboratory tests (e.g. “ref lab”). We then obtained actual laboratory results from BTRIS that were coded with these terms.
Step 2: LOINC mapping

We selected samples from these laboratory test results and designed mapping strategies according to laboratory test types. For microbiology tests, our initial focus in the study, we extracted the laboratory test information from Observation_Value_Text or Observation_Note. After a simple normalization (e.g. spacing and capitalization), we used regular expression functions to parse comment strings into four parts: 1) Test requested; 2) Test results; 3) Laboratory location; and 4) Additional information. After removing duplicate “test requested” strings, we used test names, laboratory locations, and test results for LOINC mapping.

For tests performed at a known laboratory (e.g., Mayo Medical Laboratories), we reviewed the laboratory’s Web site to obtain test names, synonyms and LOINC codes where they were available. We used RELMA to map additional test terms to LOINC codes by matching main parts (e.g. components measured, the unit of amount of substance, timing and sample type) with available information found in test results. For example, for the test name “B. burgdorferi ELISA serum,” we obtained information from the result “IgM-Negative and IgG-Negative” and matched 4 LOINC parts: component (B. burgdorferi Ab.IgG & IgM), method (ELISA), system (serum), and Scale (Ord). We classified coding results as either “mapped”, “likely mapped (ambiguous match)” or “not mapped”.

Step 3: Developing local codes

We checked the RED for existing test terms with the corresponding LOINC codes; where they did not exist, we created new local test terms and annotated them with information about the outside laboratory tests.

Step 4: Developing a reference table for fuzzy lookup transformation

Based on the results of Steps 2 and 3, we created a reference table to associate the text results of outside laboratory tests to the mapped LOINC codes and local codes, which will be used for fuzzy lookup transformation in an ETL process to code new data.

Results

Identified unspecific outside laboratory tests in BTRIS

We found that the majority of outside laboratory tests had names that conveyed what the tests were being performed (e.g. “Anti-Influenza Virus B IgG Antibody Serum Test by Mayo”). We found that nonspecific tests most often included the word “other” in their names. Table 1 shows the summary statistics of distinct RED concepts and patient rows for all the records associated with RED concepts for outside laboratory tests in the Event_Measurable table.

We identified a total of 27 RED concepts used to code outside unspecific laboratory tests. Among them, 4 RED concepts are associated with microbiology tests: “other microbiology test,” “other mayo clinic microbiology Test,” “other micro mayo contract laboratory test,” and “other micro AML Test.”

We extracted patient rows from the joined tables of Event_Measurable and Observation_Measurable where the value of Event_Name_CONCEPT was one of these 27 concepts. As a result, we obtained 14,082 unspecified laboratory test results that are stored in Observation_Note and Observation_Value_Text columns. We found each of the microbiology tests had a specific, consistent report pattern, including units, and reference ranges, although naturally the values and interpretations varied.

Table 1. Counts of distinct RED concepts and incidences in the event table.

<table>
<thead>
<tr>
<th>RED Concepts</th>
<th>Patient Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>All lab orders</td>
<td>8,549 89,105,930</td>
</tr>
<tr>
<td>Identified outside lab</td>
<td>881 252,361</td>
</tr>
<tr>
<td>Outside unspecific lab</td>
<td>27 32,590</td>
</tr>
</tbody>
</table>

LOINC mapping

From the “other microbiology test”, we selected 1,000 patient data rows as a convenience sample data set and extracted 342 unique laboratory test names from the comment fields. Of these, 298 test names were from 16 known laboratories (Table 2). We mapped 329 of 343 (95.9%) laboratory tests to 102 unique LOINC codes through laboratories’ Web sites and RELMA. No match was found for 14 tests because of the lack of minimal information, or a new LOINC code required (see examples in Table 3).
Table 2. LOINC mapping results.

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>Test Names</th>
<th>Website</th>
<th>RELMA</th>
<th>Unmapped</th>
<th>Unique Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mayo Medical Laboratories</td>
<td>106</td>
<td>100</td>
<td>5</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Beacon Diagnostics Laboratory</td>
<td>61</td>
<td>61</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>45</td>
<td>42</td>
<td>3</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Quest/Focus Diagnostics</td>
<td>40</td>
<td>39</td>
<td>1</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>MiraVista Diagnostics</td>
<td>28</td>
<td>28</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fungus ing Lab, University of Texas</td>
<td>11</td>
<td>11</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>11</td>
<td>11</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViroMed Laboratories</td>
<td>11</td>
<td>11</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centers for Disease Control and Prevention</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Maryland State Lab</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>National Jewish Health</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Specialty Laboratories</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Center for Anti-Infective Research, Hartford, CT</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>University Of Minnesota</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immunetics</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palo Alto Medical Foundation</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Medical Laboratories</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Examples of unmapped test terms.

<table>
<thead>
<tr>
<th>Test Requested</th>
<th>Test Results</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta-D-Glucan</td>
<td>No Content</td>
<td>Lack information to determine the test in process of code application (or mapping)</td>
</tr>
<tr>
<td>Blastomycoses Ab, EIA, S</td>
<td>Negative</td>
<td>In process. No panel for antymycobacterial susceptibility testing</td>
</tr>
<tr>
<td>Anti-mycobacterial Drug Panel</td>
<td>Ethambutol Level 0.97 Micrograms/Ml</td>
<td>The result indicates a test 'Microsporidia Molecular Identification,' that can not be inferred from 'Q&amp;P.' No match in LOINC using 'PCR.' Also, this example shows the inconsistency between lab test and result.</td>
</tr>
<tr>
<td></td>
<td>Isoniazid Level 0.94 Micrograms/Ml</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pyrazinamide Level 14.46 Micrograms/Ml</td>
<td></td>
</tr>
<tr>
<td>Ova and Parasite Identification</td>
<td>Microsporidia Detected By PCR</td>
<td></td>
</tr>
<tr>
<td>West Nile PCR</td>
<td>West Nile Virus, Eastern Equine Encephalitis Virus Saint Louis Encephalitis Virus, Lacrosse Virus Not detected by RT-PCR.</td>
<td>A full arboviral panel conducts when a West Nile is ordered.</td>
</tr>
<tr>
<td>Bordetella Pertussis Antibodies, Igg</td>
<td>B. pertussis Ab, IgG W/Reflex 2.0 U/Ml B. pertussis Ab, IgM W/Reflex 1.2 U/Ml, B. pertussis Ab, IgM Immunoblot: No IgM Antibodies A against Bordetella FHA and PT detected.</td>
<td>This is a reflex test. The naming of panels with reflex components in LOINC has not been addressed.</td>
</tr>
</tbody>
</table>

Developing local codes

We added 102 new terms to the RED when no laboratory test name had been represented in the RED, based on mapping and information from the outside laboratory.

Creating a reference table for fuzzy lookup transformation

Multiple columns are included on the reference table for fuzzy lookup transformation, including local code, local term, full text string, LOINC code and name for laboratory results. Table 4 shows the structure of the reference table with sample data.
Discussion

Data in EHRs are often encoded or coded to a level that is inadequate for re-use. Outside laboratory test results, such as those in our study, are commonly found in EHRs and are particularly troublesome, especially when attempting to merge such data with data that are more explicitly coded. In this study, we demonstrated that such tests do actually have appropriate LOINC codes and automated methods can be used to achieve mapping where neither the outside laboratory nor the receiving laboratory system are able to do so.

LOINC mapping. The LOINC code is an intermediate between local and outside laboratory tests. Our mapping is based on textual comments reported by the outside laboratory as results that actually include test names and results, which are different from mapping local laboratory tests based on the names from data dictionary. Our mapping strategy is a specific-general approach. We choose a specific LOINC code for the laboratory result with multi-part matching. Considering test results with less information, we prefer a relatively general code that can cover laboratory results with similar reporting patterns. The laboratory results from well-known laboratories use similar report formats for specific tests, so that the mapping results can be shareable within any laboratory system that obtains results from these laboratories. Our general mapping strategy produces a high success rate; however, we do not consider the possibilities of multiple LOINC codes and the influences of specific methods used in the laboratory tests. Although our mappings appear correct, we have not conducted a formal evaluation of their accuracy.

Local coding. We need to add new terms or codes into the list of laboratory tests, if these terms are not in an EHR system. In clinical data repositories within the i2b2 model, the i2b2 ontology management cell manages new test names under accurate concept paths. The RED concept structure allows the representation of outside laboratory test information in "synonym" properties and their associated attributes as local data sources. However, when we add as many new terms as we can identify from unspecified outside laboratory test results, we bear the burden of the need for additional concepts and an efficient method to manage the concepts.

Fuzzy lookup transformation. Microsoft SQL server integration services (SSIS) provides a platform for ETL processes that supports routines for outside laboratory test results transformation into the daily work flow. In BTRIS, we will apply the featured fuzzy lookup function in the ETL processes to identify unspecified laboratory results to match them with previously encountered test result (perhaps differing only by numeric result or punctuation). The input data source is the file with original outside laboratory results. The program will produce similarity scores for each sample in the reference table. The input data will be added to the output file with the specific laboratory test name with the highest similarity score above a given threshold. So far, we have worked on sample data retrieved from BTRIS. Additional experiments on new incoming laboratory data are needed.

Conclusion

We identified the meanings of outside unspecified laboratory tests results using medical standards and developed an outside laboratory test reference table for use in fuzzy lookup transformation processes. This study suggests that a
modest effort can lead to improved coding of these outside nonspecific laboratory data, such that we do not have to settle for having a portion of the patient’s records be unusable but rather can bring these additional data to bear on data re-use tasks. The heterogeneous and dynamic nature of EHR data reminds us of the challenges in the implementation of standards.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and Lister Hill National Center for Biomedical Communications (LHNCBC). This research was also supported in part by an appointment to the NLM Research Participation Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy (DoE) and the NLM.

Disclaimer

The views and opinions of the authors expressed herein do not necessarily state or reflect those of the National Library of Medicine, National Institutes of Health or the US Department of Health and Human Services.

Competing Interests

None

References

Homophily of Vocabulary Usage: Beneficial Effects of Vocabulary Similarity on Online Health Communities Participation

Albert Park, PhD\textsuperscript{1}, Andrea L. Hartzler, PhD\textsuperscript{2}, Jina Huh, PhD\textsuperscript{3}, David W. McDonald, PhD\textsuperscript{4}, Wanda Pratt, PhD\textsuperscript{1,5}  
\textsuperscript{1}Biomedical Informatics & Medical Education; \textsuperscript{2}Human Centered Design & Engineering; \textsuperscript{3}Information School, University of Washington, Seattle, WA  
\textsuperscript{4}Group Health Research Institute, Seattle, WA  
\textsuperscript{5}Department of Media and Information, Michigan State University, East Lansing, MI

ABSTRACT

Online health communities provide popular platforms for individuals to exchange psychosocial support and form ties. Although regular active participation (i.e., posting to interact with other members) in online health communities can provide important benefits, sustained active participation remains challenging for these communities. Leveraging previous literature on homophily (i.e., “love of those who are like themselves”), we examined the relationship between vocabulary similarity (i.e., homophily of word usage) of thread posts and members’ future interaction in online health communities. We quantitatively measured vocabulary similarity by calculating, in a vector space model, cosine similarity between the original post and the first reply in 20,499 threads. Our findings across five online health communities suggest that vocabulary similarity is a significant predictor of members’ future interaction in online health communities. These findings carry practical implications for facilitating and sustaining online community participation through beneficial effects of homophily in the vocabulary of essential peer support.

INTRODUCTION

Many people use online health communities, such as WebMD and Facebook health groups, to exchange peer support and connect with others\textsuperscript{1}. Research on the benefits of online health communities highlights psychosocial benefits—such as reduced depression\textsuperscript{2,3} and stress\textsuperscript{4,5}—from active participation in online health communities. However, sustaining active participation remains a prominent challenge for online communities in general\textsuperscript{6-10} due to issues like lurking (i.e., participating without posting) and dropouts.

Sustained, active participation in online communities has been shown to positively correlate with a number of different factors. For example, receiving a response to a newcomers’ first post\textsuperscript{8}, receiving emotional support\textsuperscript{4}, obtaining a sense of community\textsuperscript{11}, and having familiarity with online interactive services (e.g., chat)\textsuperscript{6} have all shown to positively correlate with active participation or degree of effort and time spent with the community. Although these studies provide insight on how to sustain active participation in general online communities, only one study examined online health communities\textsuperscript{8}. In contrast to the typical online community, active participation in online health communities could have implications for quality of life due to the purpose of participation: exchanging health information and psychosocial support. In that study of an online health community,\textsuperscript{8} participation was measured by sign-ins, which includes passive behaviors like lurking\textsuperscript{10}. Lurkers not only gain fewer benefits than active participants\textsuperscript{4} but they also do nothing to promote community sustainability. In our study, we focus on active participation to better reflect the benefits of online health communities to members and community sustainability.

Homophily, the tendency for individuals to be attracted to others with similar characteristics such as attitude and behavior mimicry\textsuperscript{12}, is an important yet underexplored principle in studying online community participation. However, homophily is a well-established principle in the context of social network analysis\textsuperscript{13}, and it has been shown to positively correlate with credibility of authors in online health communities\textsuperscript{14}. Moreover, homophily expressed through unconscious mimicry was correlated with likelihood of liking the respondents\textsuperscript{15}. Similarly in language, verbal mimicry of function words (i.e., content-free parts of speech) has been shown to positively correlate with liking the respondents\textsuperscript{16} and positively functioning social dynamics\textsuperscript{17}. In online health communities, both function words and content words can serve as important cues for measuring homophily expressed in vocabulary usage. The function words are related to unconscious mimicry whereas the content words are related to similarity in health traits. Although homophily measured using all types of vocabulary usage—vocabulary similarity—could have effects on members’ active participation, vocabulary similarity has not been studied with respect to active participation in online health communities.
Based upon previous literature, we expect community members to appreciate responses in which respondents use similar vocabulary. Thus, we hypothesized that individuals are likely to sustain active participation if they receive replies written with similar vocabulary. We examined this issue within the context of five online health communities from WebMD.com to address the following research questions:

**(RQ1)** What is the relationship between receiving replies written using a similar vocabulary and original posters’ subsequent thread engagement?

**(RQ2)** What is the relationship between receiving replies written using a similar vocabulary in the early stage of joining the community and newcomers’ sustained community participation?

**(RQ3)** What factors other than homophily in vocabulary usage correlate with active participation in online health communities?

**METHODS**

Our overarching objective is to understand participation in online health communities, in particular the relationship between vocabulary similarity of received replies and the member’s future interaction in the community. We define **original post** as a post that starts a thread and **original poster** as the author of the original post. Similarly, we define **first reply** as the first post to reply to an original post and **respondent** as the author of the first reply. If the original poster or respondent uses multiple posts consecutively, we considered the accumulation of those posts as the original post or first reply, respectively. For example, original posters and respondents occasionally add comments in subsequent posts before any other member replies. Hence we included any supplementary posts as a part of the original post/first reply. We define **reengagement** as the behavior of original posters returning back to threads they started and having further conversation with the respondent (i.e., by posting a reply). Conversely, we defined **disengagement** as the behavior of original posters not posting a reply to that thread.

In our analysis, we restricted our focus to the first reply for two reasons. First, we wanted to pick the post with the highest chance of reaching the original poster. First replies appear for the longest time compared to other posts in the thread. Hence original posters have the longest time to read first replies. Second, systematically assessing who is responding to whom is difficult without analyzing the content of each post. For instance, the third person to post (the second replier) could be interacting with the respondent or the original poster. Those posts that are not replying back to the original post could skew the results; thus, we focused our analysis on first replies.

We reviewed common approaches from information retrieval that could be used to quantify **vocabulary similarity** that would represent homophily of vocabulary usage between original posters and respondents. We decided to use a vocabulary-based cosine similarity measurement without any feature reductions (e.g., removing common words) to quantify vocabulary similarity score. We chose to use cosine similarity because it is one of the most common and thoroughly studied measures. One advantage of cosine similarity over other text similarity measures, such as Jaccard similarity, is that cosine similarity normalizes the text length during the comparison. Thus, long first replies would not necessarily be considered to have higher number of shared words. To determine the cosine similarity between original posts and first replies, we first represent each post as vector in N-dimensional space, where N is the number of unique terms across all posts and the value is the frequency with which terms occur in that post. Cosine similarity measures the cosine of the angle between two vectors representing the posts. The resulting similarity score ranges from zero to one. A score of zero indicates no shared terms between the two posts, whereas a score of one indicates all terms and the relative proportion of the terms used are exactly equal.

**(RQ1)** What is the relationship between receiving replies written using a similar vocabulary and the original posters’ subsequent thread engagement?

To examine RQ1, we investigated whether original posters reengaged or disengaged in the threads given the vocabulary similarity score of first replies to original posts. We applied statistical tests (i.e., Pearson’s Chi squared test (X²) and Welch’s t-tests (t) for two unpaired samples with unequal variances) to determine whether original posters who received replies with higher similar vocabulary scores reengaged more often. Thus, we compared the mean vocabulary similarity score among original posters who reengaged with the mean vocabulary similarity score among original posters who disengaged.

Next, we used logistic regression to predict the likelihood of original posters reengaging in their threads given vocabulary similarity score. Logistic regression is a statistical technique for predicting dichotomous outcome variables (i.e., engagement) given one or more predictor variables (i.e., vocabulary similarity scores). Logistic
regression limits the range of outcome variables from zero to one, satisfying assumptions for dichotomous outcome. Then, we tested the overall effects of vocabulary similarity score using the Wald test.

(RQ2) What is the relationship between receiving replies written using a similar vocabulary in the early stage of joining the community and the newcomers’ sustained community participation?

We applied survival analysis to examine the relationship between newcomers receiving replies written using a similar vocabulary to their own posts in the early stage of joining the community and the newcomers’ sustained participation in the community over time. To identify newcomers, we selected members who contributed at least one original post and received at least one first reply in their newcomer stage. The threshold for the newcomer stage was defined as up to three original posts. We chose this threshold because members with less than three posts were considered lurkers, who are not yet a regularly contributing member, in a prior study10.

Survival analysis is a time duration analysis that models survival time until the failure event occurs. We define the survival object (i.e., “sustained participation”) as the period of time in which members continue to participate in the online health community. Defining survival time with respect to online participation can be difficult because the failure event cannot be as clearly defined as in other fields, such as biological and medical sciences where survival analysis has been widely used. In the context of online health communities, members can always return to the community after years of absence as long as the community is active. We adopted a definition of a failure event from Wang et al.5 to be a period of inactivity of three months without posting to the community. We considered members’ first post (i.e., either original post or replying post in threads) as the starting point of their participation in the community and their last post as the end of their participation. However, if members posted within three months of the data collection date, we considered them right censored (i.e., member who did not experience the failure event) because they might still be actively participating in the community. We calculated the survival time as the days between members’ first and last post.

(RQ3) What factors other than homophily in vocabulary usage are correlated with active participation in online health communities?

We selected a random sample of 100 original-first reply post pairs by selecting 10 pairs that reengaged and 10 pairs that disengaged in each of the five communities. We manually examined these 100 threads for other factors related to active participation. We drew on findings from previous studies to guide our content analysis. Previous studies have shown that types of social support19 sought by original posters (e.g., informational or emotional), types of social support that original posters received8, length of original posts and first replies20,21, and rhetorical elements (i.e., asking questions9) are associated with participation.

In addition, we considered coverage of information—whether replies address all of the concerns expressed in original posts. In information retrieval, cosine similarity is used to measure the similarity of two documents with respect to their subject18. Because cosine similarity can calculate homophily of vocabulary and similarity of two documents—a proxy for coverage of information—we investigate a possible correlation between information coverage and active participation to have a deeper understanding of the effects of homophily in vocabulary.

We blindly examined the effect of these factors on future interactions with respondents. Furthermore, we explored the purpose of original posters’ reengagement. The review of types of emotional support and the purpose of original posters’ reengagement followed an open coding process22, which is a method used to elicit unknown, emerging themes grounded in data. For informational support, we assessed whether the original posters were seeking information or not.

Data: Selection and Overall Characteristics

To meet our study aims, we restricted our analysis to five communities from WebMD.com for several reasons. First, we selected chronic disease-related communities. This criterion eliminated non-disease specific communities, such as parenting and baby’s first year or smoking cessation, with a possible correlation between short-term health issues and dropout rates of WebMD forums. Second, we selected highly active communities that ranked within the top 20 WebMD forums in total number of threads. This eliminated communities that could have member dropouts due to the low-activity of the community. Third, we selected communities with two or more moderators who helped as respondents. This eliminated communities that could have member dropouts due to the low-level of moderating. Lastly, we selected communities with a sufficient number of posts from members, at least 50 first replies from both members and moderators. After applying these inclusion and exclusion criteria, five communities remained eligible for subsequent analysis: (1) ADHD, (2) Diabetes, (3) Heart Disease, (4) Pain Management, and (5) Sexual Health.
communities (Table 1). We excluded two types of original posts: (1) original posts without replies and (2) original posts started by moderators. We removed these types of original posts to focus our analysis on members and their participation behavior when receiving replies. We sought review by University of Washington Institutional Review Board (IRB) and the data was exempt from review.

Table 1. Characteristics of five WebMD communities studied

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>Diabetes</th>
<th>Heart Disease</th>
<th>Pain Management</th>
<th>Sexual Health</th>
</tr>
</thead>
<tbody>
<tr>
<td># threads analyzed</td>
<td>1,655</td>
<td>4,964</td>
<td>3,368</td>
<td>3,350</td>
<td>7,162</td>
</tr>
<tr>
<td># members as original posters</td>
<td>1,484</td>
<td>2,459</td>
<td>2,817</td>
<td>2,752</td>
<td>5,766</td>
</tr>
<tr>
<td># members as respondents</td>
<td>340</td>
<td>229</td>
<td>129</td>
<td>426</td>
<td>1,238</td>
</tr>
</tbody>
</table>

RESULTS

First reply distribution

First replies in our data set were posted within a day (mean of 21 hours), and 99% were posted within a week. Although replies posted later have an increased chance of not reaching the original posters, the reengagement rate for late replies that came after a week (9%) was comparable to the entire dataset’s reengagement rate (17%). Thus, all first replies were included in the following analysis.

Results for (RQ1): What is the relationship between receiving replies written using a similar vocabulary and the original posters’ subsequent thread engagement?

As shown in Table 2, we observed zero-inflated data distribution in which frequent zero vocabulary similarity scores were detected. Zero vocabulary similarity scores often resulted from the limited terms posted by respondents in first replies (e.g., “Find another doctor” or “no comment”). Original post and first reply pairs with zero vocabulary similarity scores had mean of 14 terms (Standard Deviation (SD)=29) in the first replies whereas the pairs with non-zero vocabulary similarity scores had significantly higher mean of 129 terms (SD=134; t(800)=70.21, p<0.0001) in the first replies. Because short generic replies are common in online communities, it was important that we keep the posts with zero similarity scores. We solved the high number of zero vocabulary similarity scores problem by fitting the data into a two-part model (i.e., zero-inflated continuous data model). We then analyzed original post and first reply pairs with zero and non-zero vocabulary similarity scores separately.

In the first part of the two-part model, we compared reengagement associated with vocabulary similarity score of zero versus non-zero. We compared the reengagement rate between the zero data set and the non-zero data set, which was significant ($X^2(1, N=20,499) = 15.27, p<0.0001$). Thus, having any vocabulary similarity score is associated with significantly higher reengagement.

In the second part of the model, we applied Welch’s t-tests to the non-zero portion of data to compare reengagement and disengagement by the original poster. Overall, Welch’s t-tests showed significantly higher vocabulary similarity score for reengagement compared to disengagement by the original poster in all communities except for Heart Disease (Table 3). Table 3 also shows the percentage of threads that original posters reengage later in the thread after multiple posters have posted. This data shows a full picture of original posters’ engagement behaviors. We suspect that no difference was found in the Heart Disease community because of its overall higher rate of disengagement. In the Heart Disease community, original posters disengaged 81% of threads, which is 20% higher than the average disengagement rate of 60% in the other four communities.

Table 2. Proportions of original post-first reply pairs with zero and above zero vocabulary similarity score by type of engagement by the respondent.

<table>
<thead>
<tr>
<th></th>
<th># zero scores</th>
<th># non-zero scores</th>
<th>Total # of pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reengagement</td>
<td>51</td>
<td>4,330</td>
<td>4,381</td>
</tr>
<tr>
<td>Disengagement</td>
<td>336</td>
<td>15,782</td>
<td>16,118</td>
</tr>
</tbody>
</table>
Table 3. Comparison of mean similarity and by type of engagement, and percentage of reengagement by original posters, disengagement by original posters, and reengagement by original posters later in the thread after multiple posters have posted

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD) vocabulary similarity for reengagement</th>
<th>Mean of vocabulary similarity for disengagement</th>
<th>Comparison of vocabulary similarity score: Reengagement vs. disengagement</th>
<th>% of reengaging</th>
<th>% of disengaging</th>
<th>% of reengaging later in thread</th>
</tr>
</thead>
<tbody>
<tr>
<td>All five communities</td>
<td>0.38(0.15)</td>
<td>0.35(0.15)</td>
<td>t(6,637)=13.45 p&lt;2.2e-16</td>
<td>17%</td>
<td>63%</td>
<td>20%</td>
</tr>
<tr>
<td>ADHD</td>
<td>0.42(0.16)</td>
<td>0.38(0.15)</td>
<td>t(370)=4.07 p=5.8e-05</td>
<td>15%</td>
<td>76%</td>
<td>9%</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.34(0.15)</td>
<td>0.32(0.15)</td>
<td>t(2,186)=5.58 p=2.687e-08</td>
<td>19%</td>
<td>52%</td>
<td>30%</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>0.38(0.13)</td>
<td>0.37(0.12)</td>
<td>t(680)=1.58 p=0.11</td>
<td>14%</td>
<td>81%</td>
<td>5%</td>
</tr>
<tr>
<td>Pain Management</td>
<td>0.43(0.15)</td>
<td>0.37(0.15)</td>
<td>t(1,242)=9.61 p&lt;2.2e-16</td>
<td>19%</td>
<td>63%</td>
<td>18%</td>
</tr>
<tr>
<td>Sexual Health</td>
<td>0.39(0.16)</td>
<td>0.34(0.15)</td>
<td>t(2,351)=10.80 p&lt;2.2e-16</td>
<td>17%</td>
<td>62%</td>
<td>20%</td>
</tr>
</tbody>
</table>

Next, we used logistic regression with one predictor variable—vocabulary similarity score—to predict the likelihood of original posters reengaging in their threads. Figure 1 shows the plot of the predicted probability with 95% confidence intervals of reengagement given the vocabulary similarity score between original post and first reply. This regression model was significant (X²(1)=91.43, p=1.55e-43) with odds ratio of 4.70. Using the vocabulary similarity score, we predicted future participation with 79% accuracy in a 10-fold cross validation. The Wald test indicated that for a one unit increase in vocabulary similarity score, the odds of original posters reengaging increased by a factor of 4.7 (X²(1)=188.60, p=6.43e-43).

**Results for (RQ2): What is the relationship between receiving replies written using a similar vocabulary in the early stage of joining the community and the newcomers’ sustained community participation?**

We applied survival analysis to test the effect of receiving replies written using a similar vocabulary on members’ sustained participation in the community. We partitioned members into three equally sized groups corresponding to members exposed to replies with a “High,” “Medium,” or “Low” vocabulary similarity score. For members with more than one original and corresponding first reply, we took the average vocabulary similarity score among the first three original and corresponding first replies. Low vocabulary similarity scores ranged from 0 to 0.28, Medium scores ranged from greater than 0.28 to 0.41; and high scores ranged from greater than 0.41 to 0.83, which was the highest vocabulary similarity score in our dataset. Examples of high and low replies are shown in Results section for RQ3.

![Figure 1. Predicted probabilities of reengagement graph with 95% confidence intervals.](image-url)
Figure 2 illustrates the effect of receiving replies written using a similar vocabulary on members’ sustained active participation. Members in the High group were most likely to stay active in the community, followed by members in the Medium group, followed by members in the Low group as least likely to stay active. These differences were sustained between the high and low groups for at least 300 days.

Results of two survival models are shown in Table 4. Model 1 reports the effects of the covariates. For instance, the hazard ratio of 0.75 for the total number of original posts indicates that those who initiate threads one standard deviation more have a 34% (i.e., (1/0.75) – 100%) higher survival rate. Similarly, Model 2 shows that members who received replies with a vocabulary similarity score of one standard deviation higher have a 5% (i.e., (1/0.95) – 100%) higher survival rate when controlling for covariates.

The hazard ratio indicates the odds of members dropping out of the community (encountering the failure event). We also considered a number of covariates and their relationship to sustained participation in two survival models. We selected covariates representative of each member’s intrinsic characteristics (e.g., sociable) as well as the amount of participation in the community. These variables include the total number of posts, total number of first replies provided, total number of first replies received, and total number of original threads. We normalized variables (i.e., (observation – mean)/standard deviation) to show predicted change in odds for a unit increase in the predictor.

**Table 4. Survival analysis showing influence of covariates in two models**

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Model 1 Hazard Ratio</th>
<th>Model 1 Standard Error</th>
<th>Model 2 Hazard Ratio</th>
<th>Model 2 Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of posts</td>
<td>0.91**</td>
<td>0.034</td>
<td>0.92**</td>
<td>0.034</td>
</tr>
<tr>
<td>Total number of first replies provided</td>
<td>0.92**</td>
<td>0.032</td>
<td>0.92**</td>
<td>0.031</td>
</tr>
<tr>
<td>Total number of first replies received</td>
<td>0.83*</td>
<td>0.091</td>
<td>0.84</td>
<td>0.091</td>
</tr>
<tr>
<td>Total number of original threads</td>
<td>0.75**</td>
<td>0.095</td>
<td>0.74**</td>
<td>0.094</td>
</tr>
<tr>
<td>Vocabulary similarity scores</td>
<td></td>
<td></td>
<td>0.95***</td>
<td>0.008</td>
</tr>
</tbody>
</table>

***: p<0.001, **: p<0.01, *: p<0.05

Results for (RQ3): What factors other than homophily in vocabulary usage are correlated with active participation in online health communities?

Without any knowledge of their vocabulary similarity scores or reengagement status, we manually categorized original post and first reply pairs into three groups: high, medium, and low coverage groups. We categorized pairs with first replies that addressed all of the concerns expressed in original posts as high coverage, first replies that addressed some concerns as medium coverage; and first replies that did not address any concerns as low coverage. We then examined how well the vocabulary similarity measures performed compared to manual categorization. High coverage group compared to low coverage show significantly higher vocabulary similarity scores (t(19)=2.58, p=0.02) (Table 5). However, the difference between high coverage group and medium coverage group (t(20)=0.34, p=0.74) as well as the difference

**Table 5. A comparison among subjective and vocabulary similarity scores**

<table>
<thead>
<tr>
<th>Vocabulary similarity scores</th>
<th>Mean of high coverage (SD)</th>
<th>Mean of medium coverage (SD)</th>
<th>Mean of low coverage (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.40 (0.14)</td>
<td>0.39 (0.18)</td>
<td>0.29 (0.16)</td>
</tr>
</tbody>
</table>
between medium coverage group and low coverage group ($t(29)=1.63, p=0.11$) were not significant.

In the 50 reengaging pairs, we found that 86% provided high coverage, 8% provided medium coverage, and 6% provided low coverage. In contrast, of the 50 disengaging pairs, 52% provided high coverage, 24% provided medium coverage, and 24% provided low coverage. Comparison of high, medium, and low information coverage showed reengaging had a significantly higher proportion of high coverage replies ($X^2(1, N=100)=11.97, p=0.0005$). As well, disengaging had a significantly higher ($X^2(1, N=100)=5.02, p=0.03$) proportion of low coverage replies. However, medium coverage did not differ significantly ($X^2(1, N=100)=3.65, p=0.06$).

Similarly, when we compared pairs associated with reengagement and with disengagement, we found a noticeable difference in the length of the posts measured by number of words. The pairs associated with reengagement used more words in both original posts and first replies. On average in the reengaging pairs, original posters used 198 words (SD=152) while respondents used 181 words (SD=150). In contrast, in the disengaging pairs, original posters used 122 words (SD=116) while respondents used 137 words (SD=142). The difference between reengaging and disengaging pairs was significant in original posts ($t(92)=2.78, p=0.006$) but not significant in first replies ($t(98)=1.48, p=0.14$).

We also found differences in emotional support: whether respondents indicated an aspect of empathizing with or helping original posters. We found two themes of emotional support: acknowledgement of members’ experience of difficulty (e.g., “I know exactly how you feel [...] I’m in the same boat”) and encouragement for the current situation (e.g., “You’ve got a great attitude, and will do well”). In reengaging pairs, 30% of respondents acknowledged the difficulty of the original poster’s situation and 34% of respondents encouraged original posters. Conversely, in disengaging pairs, 18% and 28% of respondents acknowledged their difficulty and provided encouragement, respectively. However, differences in these propositions were not significant (Acknowledgement: $X^2(1, N=100)=1.37, p=0.24$; Encouragement: $X^2(1, N=100)=0.19, p=0.67$).

The overall exchange of emotional support was less frequent than informational support; this fits well with what original posters were seeking. Overall, 89% of the original posters asked for new information, while only 24% of the original posters showed any signs of requesting emotional support (e.g., “can’t take more” or “desperate!”). Exchanges of emotional and informational support were not mutually exclusive as some original posters sought both.

We also examined the effects of respondents asking questions to original posters in first replies. In reengagement, 32% of respondents asked a question in their reply, while only 20% of respondents asked a question in disengagement, however, difference in these proportions were not significant ($X^2(1, N=100)=1.30, p=0.25$). Similarly difference in the proportions of providing informational support was not significant ($X^2(1, N=100)=0.07, p=0.79$) between reengagement and disengagement. In reengagement, 84% provided informational support, while 80% provided informational support in disengagement.

In our qualitative analysis, we identified four themes for the purpose of original posters’ reengagement: (1) providing more information on their situation (82%), (2) thanking the original posters (62%), (3) asking more questions (20%), and (4) getting defensive (8%). These behaviors were not mutually exclusive.

In our sample, reengaging and disengaging pairs varied in the degree of information coverage. First replies that covered more aspects of the original post received higher vocabulary similarity scores than those that covered fewer aspects. The following is an example of a reengaging pair that received a relatively high vocabulary similarity score of 0.66.

**Original poster A:** “I’d much rather ask others that have a condition I may have... I’ve been pretty hyper all my life [...] and have been bouncing around the idea in my head that I may have some sort of ADHD. [...]I’ve noticed I can focus more when I am very tired. [...]”

**Respondent to A:** “You are describing Adult ADHD. You should make an appointment with a psychiatrist who was experience with adult ADHD. [...] My mind is usually all over the place bouncing from topic to topic [...] Things aren't like that now. [...] Go. Get tested. [...] If you do have ADHD, he may or may not prescribe medication. [...]You have nothing to loose as long as your honest with your doctors. Good luck. [...]”

Conversely, the following is an example of a disengaging pair that received relatively low vocabulary similarity score of 0.14.
Original poster_B: “been having muscle ache between upper left chest (near armpit) through shoulder [...] Did I strain a muscle/group of muscles, and what's best - heat, NSAIDs, chiropractic?”

Respondent_to_B: “Hello. No way for any of us on an internet message board to know what is causing your symptoms. You will need to see your doctor for an evaluation and treatment plan [...]”

In both examples, the respondents advocate seeking professional help. However, only in the first example, did respondent_to_A provide their experience and perspective. During this process, respondent_to_A covered a number of issues raised by original poster_A while using more shared vocabulary. This conversational pair resulted in a relatively high vocabulary similarity score and elicited original poster_A to reengage in further conversation. In the second example, respondent_to_B does not address the concerns original poster_B raised. This resulted in using less shared vocabulary with original poster_B and a relatively low vocabulary similarity score.

As mentioned earlier, other factors also correlate with original posters’ engagement. Respondent_to_A acknowledges the difficulty and encourages original poster_A while using more words (521 words) than respondent_to_B (34 words). Conversely respondent_to_B is succinct and does not provide any emotional support while neither respondent asked questions. Although a combination of many factors can influence engagement of the original poster, in our manual analysis receiving replies with more shared vocabulary appears to be associated with reengagement, which supports our quantitative analysis.

DISCUSSION AND FUTURE WORK

Prior research shows that sustaining active participation presents a prominent challenge for online communities6–10. In this paper, we showed the importance of homophily expressed through shared vocabulary associated with members’ ongoing engagement in online health communities. Members who received replies that contained more shared vocabulary with their own posts tended to continue their conversations with respondents.

Additionally, we created prediction models that estimate the likelihood that original posters will reengage with respondents. Although many factors can contribute to members disengaging from conversation, our logistic regressions showed that vocabulary similarity predicts future participation with the respondents. Similar prediction models could be one solution for the challenge of sustaining active participation. For instance, our prediction model could allow moderators to identify members who are most likely to disengage. This added knowledge could enable moderators to provide replies that encourage reengagement. Moreover, we discovered that receiving replies written in a similar vocabulary in the early stages of joining the community predicts long-term active participation within the community. One solution for sustaining newcomers’ participation in the community is to encourage community members to abide by a set of guidelines, which reflect ideal community member interactions with newcomers. Furthermore, measuring vocabulary similarity can be a basis for automatically alerting members when they deviate from posting replies that encourage reengagement. In contrast with targeting specific members to encourage reengagement, the environment of the community as a whole could be improved by filtering spam or abusive content through comparing terms with the common vocabulary of the community.

In our manual analysis, we found that a combination of many different factors could influence original posters’ engagement. Other factors, such as exposure to higher degree of information coverage and higher word counts by original poster20,21 were associated with reengagement in our data. All other factors had higher occurrences in reengagement, but they are not found significant. We suspect this is due to small sample size of 100 in our manual analysis. Investigating which factors had the biggest impact on original posters’ engagement is an important question for future work. However, measuring vocabulary similarity of first reply to the original post is a relatively easy and robust technique that seems to measure an important marker for member reengagement.

Furthermore, the vocabulary similarity of the first reply might not be the sole factor of member engagement with respondents. For instance, other components of life can influence online health community participation. Original posters could have gained the needed information through other sources and did not check back with the community. A serious medical crisis could prevent original posters from checking back with the community too, for example. Still, the consistent statistical results from examining the relationship between vocabulary similarity and participation show that homophily of vocabulary provides an important marker for member reengagement.

Our survival analysis examines members’ early stages of joining the community and uses a threshold of the first three replying posts that members received based on previous literature on lurking10. Understanding the transition point of members’ participation can provide a deeper understanding of how to sustain active participation in online communities.22,23
communities. Also, our analysis does not answer whether inactive members remained lurkers or completely dropped from the community. An analysis of these two different types of inactive members could further extend our findings.

Although overall our analyses showed consistent results in a diverse group of online health communities, we acknowledge that our large sample size could have inflated the significance levels or otherwise skewed results and raises questions to the practical significance of our findings. Also, how much vocabulary similarity is needed for a meaningful increase in reengagement remains an open question. Further investigation, such as surveys or interviews, analyzing members’ satisfaction and understandability of health information correlated to vocabulary similarity could further explore the significance of our findings.

In future work, we plan to investigate the correlation between actual users’ perceived qualities of replies and participation to further examine the challenge of sustaining participation in online health communities. Understanding these relationships could provide a more complete view of how to sustain participation in online health communities. The significance of this study goes beyond predicting members’ behavior in online health communities. For instance, our findings could generalize in non-health online communities and our automatic approach to analyze computer-mediated communication (CMC) could be applied to other CMC studies. Furthermore, our study showed a potential method to elucidate the process of forming social bonds through CMC in online communities.

CONCLUSION

We provide new insights regarding sustaining online health community participation through systematic analyses of five WebMD online health communities. Our findings suggest that homophily—the vocabulary similarity between members’ posts—plays a crucial role in sustained engagement in online health community. We provide new insights into how vocabulary similarity affects active participation in online communities. Furthermore, vocabulary similarity calculated with cosine similarity shows promising results in measuring the coverage of information in replies. Based on these insights, moderators, online community creators, and online community participants could tailor replies to encourage sustained, active participation by members. Findings from this study can improve member experience in difficult situations when online health communities provide essential support.

ACKNOWLEDGEMENTS

This work was funded by NSF SHB 1117187 and NIH-NLM # 5T15LM007442-10 BHI Training Program.

REFERENCES

1. Fox S, Duggan M. Health online 2013: 35% of U.S. adults have gone online to figure out a medical condition; of these, half followed up with a visit to a medical professional. Health (Irvine Calif) [Internet]. 2013;1–55. Available from: http://www.webcitation.org/6Y5K9a3JH


Approaches to Supporting the Analysis of Historical Medication Datasets with RxNorm

Lee B. Peters, M.S., Olivier Bodenreider, M.D., PhD
National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Abstract

Objective: To investigate approaches to supporting the analysis of historical medication datasets with RxNorm. Methods: We created two sets of National Drug Codes (NDCs). One is based on historical NDCs harvested from versions of RxNorm from 2007 to present. The other comprises all sources of NDCs in the current release of RxNorm, including proprietary sources. We evaluated these two resources against four sets of NDCs obtained from various sources. Results: In two historical medication datasets, 14-19% of the NDCs were obsolete, but 91-96% of these obsolete NDCs could be recovered and mapped to active drug concepts. Conclusion: Adding historical data significantly increases NDC mapping to active RxNorm drugs. A service for mapping historical NDC datasets leveraging RxNorm was added to the RxNorm API and is available at https://rxnav.nlm.nih.gov/.

Introduction

Many electronic medical systems identify a drug product by using a National Drug Code (NDC). For example, NDCs are used for identifying prescription drugs in the Medicare “Part D” database, as well as by many pharmacies, pharmacy benefit managers and health insurance companies [1-10]. NDCs represent not only the product’s characteristics (dosage strength and form), but also manufacturer and packaging information. Because of their specificity, NDCs tend to be less stable identifiers compared to other drug vocabularies, such as RxNorm. For example, NDCs can become obsolete not only due to the product discontinuation for the usual safety reasons, but also due to discontinuation by a manufacturer for business reasons, or due to changes in packaging (e.g., pack size). In some cases, the drug may still be produced (with the same identifier in RxNorm), but by a different manufacturer or with a different pack size (i.e., under a different NDC).

Because NDCs are widely used drug identifiers, there is a great need and interest in mapping the NDC for a drug product into a standardized RxNorm name for use in electronic medical systems. In fact, the most used function in the RxNorm API is findRxcuiById, used to map a variety of drug identifiers to RxNorm, and NDC is the type of identifier most often converted to RxNorm with our API. More specifically, our API received 155 million findRxcuiById requests in 2014, 123 million (79%) of which specify NDCs.

In addition to its own identifiers, RxNorm maintains a collection of curated, up-to-date NDCs and therefore supports the mapping of current NDCs to RxNorm. However, in addition to mapping NDCs to RxNorm for current datasets or transactions from health information networks, researchers have shown interest in analyzing historical medication datasets. One such dataset is the Medicare “Part D” dataset. While RxNorm maintains history information for its own identifiers, it does not keep track of obsolete NDCs and their connection to the drugs they referred to is lost. Similarly, the NDC database maintained by the FDA only contains currently valid NDCs.

Our objective is to investigate approaches to supporting the analysis of historical medication datasets with RxNorm. More specifically, we explore NDCs collected from earlier versions of RxNorm and NDCs provided by drug vocabularies integrated in RxNorm (but not curated by RxNorm), in their coverage of several large datasets of NDCs collected from various sources and time periods. Ultimately, our goal is to support the development of a new API function for mapping legacy NDCs to RxNorm, informed by the findings of this investigation.

Background

National Drug Code (NDC). The NDC is a universal product identifier for human drugs in the United States. The Drug Listing Act of 1972 requires registered drug establishments to provide the Food and Drug Administration (FDA) with a current list of all drugs manufactured, prepared, propagated, compounded, or processed by it for commercial distribution. Drug products are identified and reported using the NDC.

The NDC is represented by a unique 10-digit, 3-segment number. The first segment of the NDC identifies the labeler (manufacturer, distributor or re-packerager). The second segment is the product code, which identifies the specific strength, dosage form (i.e., capsule, tablet, liquid) and formulation of a drug for a specific manufacturer.
The third segment is the package code, which identifies package sizes and types. The first segment (labeler code) is assigned by the FDA and the second and third segments are provided by the labeler.

Examples of 10-digit, 3-segment NDCs:
- 54868-1048-1 (5-4-1 format)
- 55111-476-79 (5-3-2 format)
- 0179-0111-70 (4-4-2 format)

Many systems, including RxNorm, convert these forms into an 11-digit NDC derivative, which pads the labeler, product, or package code segments of the NDC with leading zeroes wherever they are needed to create a 5-4-2 format without the dashes. Examples of conversion:
- 54868-1048-1 => 54868104801
- 55111-476-79 => 55111047679
- 0179-0111-70 => 00179011170

RxNorm is a standardized nomenclature for medications produced and maintained by the U.S. National Library of Medicine (NLM) in cooperation with proprietary vendors [11]. RxNorm concepts are linked by NLM to multiple drug identifiers for each of the commercially available drug databases within the UMLS® Metathesaurus®. In addition to integrating names from existing drug vocabularies, RxNorm creates standard names for clinical drugs. RxNorm also contains NDC codes in the 11-digit NDC derivative format described above. The NDCs curated by RxNorm are derived from two terminologies – DailyMed and First Data Bank. Some other source vocabularies also contribute NDCs to RxNorm. However, these additional NDCs are not curated by RxNorm. RxNorm is updated monthly and each version only contains active NDCs, i.e., those NDCs which reflect the current state of availability of drug products. With each new version of RxNorm, some are added, while others that have become obsolete are removed. RxNorm does not keep track of obsolete NDCs.

The RxNorm API [12] provides functionality to access the RxNorm dataset, including mapping from NDCs to obtain the RxNorm concept identifier (RxCUI). It accepts the NDC in the 10-digit, 3-segment sequence or as the 11-digit derivative. Only active NDCs (curated by RxNorm in the latest release) can be mapped to RxNorm concepts.

Related work. Hanna et al developed a historical NDC dataset to use in the Drug Ontology [13]. While our objective is in part similar to theirs, their main goal was to harvest historical NDCs. In contrast, the specific contribution of this work is twofold. First we investigate the enrichment of a reference set of NDC not only with historical NDCs, but also with NDCs from other drug information sources. More importantly, we provide a comprehensive evaluation of the impact of using a richer set of NDCs by measuring the results on several large sets of NDCs from various sources and time periods.

Methods

Our investigation of approaches to supporting the analysis of historical medication datasets with RxNorm can be summarized as follows. First we describe two approaches to enriching RxNorm with additional NDCs:

1. Collect curated NDCs from earlier versions of RxNorm,
2. Collect NDCs from all drug vocabularies in the latest release of RxNorm (curated or not by RxNorm).

We then evaluate these two sets of NDCs in their coverage of large datasets of NDCs collected from various sources and time periods.

Enriching RxNorm with additional NDCs

To enrich RxNorm with additional NDCs, we need to acquire NDCs from some source and to associate each NDC with a valid RxCUI in the current version of RxNorm.

Collect curated NDCs from earlier versions of RxNorm. As mentioned above, the RxNorm dataset is restricted to currently valid NDCs and does not contain historical information regarding legacy NDCs. To create an NDC
history, we retrieved the monthly releases of the RxNorm dataset starting in July 2007 through the March 2015 release to track all the curated NDCs active at any time during this time period. For each NDC, the start and end times of the period of activity were recorded, as well as the concept identifier in RxNorm with which it was associated. For some NDCs, the RxNorm concept identifier originally associated with the NDC became obsolete and was remapped. For example, NDC 00002036303 (Darvocet-N tablets) was originally linked to RxCUI = 687241, but that concept was later mapped to 849692. We refer to this set of NDCs and related information extracted from historical RxNorm versions as the History data.

Collect NDCs from all the drug vocabularies integrated in RxNorm. There exist in the RxNorm dataset NDCs which are provided by drug vocabularies integrated in RxNorm, but not curated by RxNorm. These NDCs are not part of the active set of NDCs and are not retrieved by the RxNorm API. The reasons for these NDCs not being in the active set could be that they deal with products, such as needles or syringes or that they represent experimental or unapproved drugs by the FDA. Moreover, these NDCs may come from proprietary sources and may not be publicly available. In this investigation, we looked at the March 2015 release of the RxNorm dataset and harvested all the NDCs from all the drug vocabularies. We refer to this set of NDCs and related information extracted as the All Sources data.

Associating NDCs with active RxCUIs

To determine if an NDC is active, we used the RxNorm API method findRxcuiById. For active NDCs, the RxNorm concept identifier (RxCUI) is identified. For obsolete NDCs (NDCs that were once active but are no longer active), the last active time period is identified along with the last known RxCUI from the historical data. We use the RxNorm API to determine if the RxCUI associated with the NDC is active, has been remapped, or is inactive. For example, the obsolete NDC 0018202833 (Bacitracin Ointment) was last mapped to RXCUI=308509 but that RXCUI was later remapped to 1366116. NDCs which are not contained in historical data but are contained in All Sources (which we call “Alien” NDCs) have an RxCUI associated with them, and we use the RxNorm API to determine if that RxCUI represents an active RxNorm drug product.

Based on the presence of an NDC in the various sources, we determine the status of an NDC in the following way.

- **Active.** The NDC is currently recognized by RxNorm as an active drug code (i.e., is part of the curated NDCs from the latest release of RxNorm). All active NDCs are associated with an active RxCUI by definition, and can be found by using the RxNorm API. All active NDCs are also contained in both the History data and the All Sources data.

- **Obsolete.** The NDC is no longer active, but was in the past (i.e., was part of the curated NDCs from some earlier version of RxNorm). Some obsolete NDCs can be associated with an active RxCUI (e.g., if the RxCUI to which they were originally associated is still active or can be remapped to an active RxCUI). These NDCs are found in the History data.

- **Alien.** The NDC is not recognized by RxNorm as an active drug code, nor has it been in the past, but it is currently contained in at least one drug vocabulary. This indicates the possibility of an out of scope drug product. Most of these NDCs are not associated with an active RxNorm concept. These NDCs can be found in the All Sources data (and not in the History data).

- **Unknown.** The NDC is not found in either the History data or All Sources data.

The procedure for determining the NDC status goes through the steps listed above in order and stops when the NDC is identified. Once the NDC is identified from one of the datasets, we use the procedure outlined above to find the active RxCUI (if it exists) associated with the NDC. In order to facilitate the evaluation, we created an application for looking up the NDCs and their status from a database.

Evaluation

In order to evaluate the practical benefit of using additional NDC sources to the analysis of medication datasets, we acquired several large datasets of NDCs collected from various sources and time periods. We performed a quantitative evaluation of the coverage each dataset. Additionally, we performed a qualitative analysis of the NDCs from one of these sources for which no mapping to RxNorm could be found.
Quantitative evaluation. To test the NDC status function, we used sets of NDCs from three distinct sources:

- **Medicare NDCs.** This dataset came from a random sample of Medicare Part D patients who enrolled in 2009. The set contains 27,186 unique NDCs prescribed to these patients in 2011, as well as the frequency of prescription for each NDC.

- **Private insurance NDCs.** This dataset came from a large private health insurance group and corresponds to NDCs collected during the period January 2010 to May 2014. The set contains 51,490 unique NDCs.

- **API log file NDCs.** We took the NDCs specified in the RxNorm API calls to `findRxuiByld` for a single month (January 2015) from the API log files. This API call allows the user to find the RxNorm concept associated with the user specified NDC. We removed any input that was not in the 10-digit, 3-segment format or the 11-digit derivative format and converted all to the 11-digit derivative format. The resulting dataset contained 372,705 unique NDC entries.

- **FDA NDC list.** A reference list of approved NDCs for drug products exists from the FDA. We downloaded this list for http://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm in our investigation to check against the NDCs in the History and All Sources data. The FDA list contained 167,748 NDCs.

Qualitative evaluation. One of the authors (OB), a physician, reviewed all the NDCs from the Medicare dataset for which no mapping to an active RxNorm drug could be found. This evaluation was made possible by the fact that this source provided a generic drug name for each NDC. In practice, we performed a manual review assisted by the use of regular expressions to capture frequently occurring words, corresponding to five major categories:

- Supplies (e.g., needle, syringe, lancet)
- Vitamins and dietary supplements (e.g., hyoscyamine, ferrous sulfate, carotene)
- Cold medicine (e.g., pseudoephedrine, methorphan, menthol)
- Other kinds of over-the-counter drugs (e.g., fluoride, glycerin, ointment)
- Potential prescription drugs (e.g., furosemide, insulin, nifedipine)

![Number of RxNorm NDCs](image)

Figure 1. Active NDCs in RxNorm
Results

*Enriching RxNorm with additional NDCs*

**Collect curated NDCs from earlier versions of RxNorm.** The number of unique NDCs in the History data totaled 445,039. As shown in Figure 1, the number of active NDCs in each monthly version RxNorm fluctuates over time, reflecting not only addition of new drugs, but also curation efforts to eliminate obsolete or unreliable NDCs. In recent years, the number of NDCs curated by RxNorm is about 200,000. The History data includes 418,287 NDCs which can be linked to active RxNorm concepts, which includes 214,876 NDCs not already covered by (active) RxNorm NDCs.

**Collect NDCs from all the drug vocabularies integrated in RxNorm.** The NDCs from the All Sources data include 607,451 NDCs. There are 178,284 (alien) NDCs not covered by the History data. Of these alien NDCs, only 6,485 are linked to active RxNorm concepts.

**Overlap between History NDCs and All Sources NDCs.** Figure 2 shows the overlap of the NDCs between the two datasets. The numbers in parentheses indicate the number of NDCs that are linked to active RxNorm concepts. There are 623,323 unique NDCs contained in the extended set of NDCs composed of the union of the History and All Sources sets, of which 424,772 (68%) are linked to active RxNorm concepts. The History data contains 71% of the total NDCs and 98% of the total NDCs which are linked to active RxNorm concepts. The All Sources data contains 97% of the total NDCs and 97% of the total NDCs which are linked to active RxNorm concepts.

![Figure 2. NDCs in History and All Sources](image)

(Numbers in parentheses denote NDCs linked to active RxNorm concepts)

*Evaluation*

**Medicare NDCs**

The results returned by the NDC status function for the Medicare NDC dataset are shown below (Table 1). Of the 27,186 NDCs in this dataset, 26,528 (97.6%) could be recognized when using the extended set of NDCs (“active”, “obsolete” or “alien” status), leaving only 658 (2.4%) unknown NDCs, compared to 22% unknown NDCs when only using the NDCs from the current release (“active” status). Moreover, 25,220 of all the NDCs could be linked to active RxNorm concepts, of which 24,924 (98.8%) were present in the History dataset (“active” or “obsolete” status).
Table 1. NDC status for the Medicare NDC dataset

<table>
<thead>
<tr>
<th>NDC status</th>
<th>Total # of NDCs</th>
<th># of NDCs linked to active concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>21281</td>
<td>21281</td>
</tr>
<tr>
<td>Obsolete</td>
<td>3806</td>
<td>3643</td>
</tr>
<tr>
<td>Alien</td>
<td>1441</td>
<td>296</td>
</tr>
<tr>
<td>Unknown</td>
<td>658</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>27186</td>
<td>25220</td>
</tr>
</tbody>
</table>

Private insurance NDCs
The results returned by the NDC status routine for the private insurance NDC dataset are shown below (Table 2). Of the 51,490 NDCs in this dataset, 49,767 (96.7%) could be recognized when using the extended set of NDCs (“active”, “obsolete” or “alien” status), leaving only 1723 (3.3%) unknown NDCs, compared to 33% unknown NDCs when only using the NDCs from the current release (“active” status). Moreover, 43,713 of all the NDCs could be linked to active RxNorm concepts, of which 43,359 (99.2%) were present in the History dataset (“active” or “obsolete” status).

Table 2. NDC status for the private insurance NDC dataset

<table>
<thead>
<tr>
<th>NDC status</th>
<th>Total # of NDCs</th>
<th># of NDCs linked to active concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>34529</td>
<td>34529</td>
</tr>
<tr>
<td>Obsolete</td>
<td>9724</td>
<td>8830</td>
</tr>
<tr>
<td>Alien</td>
<td>5514</td>
<td>354</td>
</tr>
<tr>
<td>Unknown</td>
<td>1723</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>51490</td>
<td>43713</td>
</tr>
</tbody>
</table>

API log file NDCs
The results returned by the NDC status routine for the API log file NDC dataset are shown below (Table 3). Of the 372,705 NDCs in this dataset, 299,230 (80.3%) could be recognized when using the extended set of NDCs (“active”, “obsolete” or “alien” status), leaving “only” 73,475 (19.7%) unknown NDCs, compared to 63% unknown NDCs when only using the NDCs from the current release (“active” status). Moreover, 222,393 of all the NDCs could be linked to active RxNorm concepts, of which 218,216 (98.1%) were present in the History dataset (“active” or “obsolete” status). The high percentage of unknown NDCs in this dataset (compared to historical medication datasets) could be explained by the fact that these NDCs came from many users, whose intentions are unclear and whose sources may have included drugs and medical products that are out of scope for RxNorm. This dataset also has a much larger proportion of Alien NDCs.

Table 3. NDC status for the API log file NDC dataset

<table>
<thead>
<tr>
<th>NDC status</th>
<th>Total # of NDCs</th>
<th># of NDCs linked to active concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>137174</td>
<td>137174</td>
</tr>
<tr>
<td>Obsolete</td>
<td>94476</td>
<td>81042</td>
</tr>
<tr>
<td>Alien</td>
<td>67580</td>
<td>4177</td>
</tr>
<tr>
<td>Unknown</td>
<td>73475</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>372705</td>
<td>222393</td>
</tr>
</tbody>
</table>
FDA NDCs
We compared the NDCs in the History and All Sources data to an external source, the FDA NDC list. When we examined the coverage of these NDCs in the FDA list, we found that 99.8% of the FDA NDCs were contained in either the History or All Sources data. On examination of the missing FDA NDCs, we found several were not valid NDCs (they contained letters) and several others we examined were for drug products that are out of scope for RxNorm.

Qualitative evaluation. We reviewed a total of 1966 unique NDCs (from the Medicare dataset) with no mapping to RxNorm, corresponding to 313,659 prescriptions. As shown in Table 4, the overwhelming majority (82%) of these prescriptions correspond to supplies. While there are quite a number of potential prescription drugs among the list, these NDCs represent a minute proportion of the entire Medicare dataset.

Table 4. Categorization of the NDCs with no mapping found to RxNorm

<table>
<thead>
<tr>
<th>Category</th>
<th># of NDCs</th>
<th># prescriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplies</td>
<td>910</td>
<td>258,218</td>
</tr>
<tr>
<td>Vitamins and dietary supplements</td>
<td>673</td>
<td>19,825</td>
</tr>
<tr>
<td>Cold medicine</td>
<td>97</td>
<td>7504</td>
</tr>
<tr>
<td>Other kinds of OTC drugs</td>
<td>162</td>
<td>784</td>
</tr>
<tr>
<td>Potential prescription drugs</td>
<td>124</td>
<td>27,328</td>
</tr>
<tr>
<td>Total</td>
<td>1,966</td>
<td>313,659</td>
</tr>
</tbody>
</table>

Discussion
Findings. The results indicate there is much to be gained from the use of historical NDC data to allow obsolete NDCs to be linked to active RxNorm concepts. In the Medicare and private insurance datasets, 14-19% of the NDCs were obsolete, but 91-96% of these obsolete NDCs could be linked to an active RxNorm concept using the History data. In contrast, while the use of all sources of NDCs in RxNorm made it possible to recognize many additional NDCs, relatively few of these Alien NDCs were linked to active RxNorm concepts and the Alien data contributed very little in additional mapping to RxNorm concepts.

Comparison with the FDA list showed that the extensive coverage of the FDA NDCs eliminates the need for it to be included in our service that identifies NDCs.

Additionally, the qualitative analysis done on a set of NDCs without mappings to active RxNorm concepts indicate most of these NDCs correspond to entities other than drugs, mostly supplies.

Application. This investigation clearly demonstrates that the analysis of historical medication datasets can benefit from exploiting sources of NDCs beyond the active NDCs (i.e., the NDCs curated by RxNorm present in the latest release). Namely, we showed that a large number of obsolete NDCs in the History data can be linked to an active RxNorm drug. Moreover, additional recognition of NDCs is supported by the All Sources data.

Until recently, the RxNorm API only supported the mapping of active NDCs to RxNorm concepts. The findings of this investigation informed the design of a new API function, getNDCStatus, to support the mapping of historical NDCs to RxNorm. This function explores the History dataset and returns all RxNorm concepts ever associated with a given NDC, along with the period when this association was active. For example, the API indicates that the obsolete NDC 00364666854 was associated with the RxNorm concept 312656 (Promazine 50 MG/ML Injectable Solution) between June 2007 and January 2011. When an NDC is associated with several RxNorm concepts at different time periods, users can select the RxNorm concept corresponding to the date of prescription, if known. The function also indicates when an NDC is recognized as “alien” (i.e., is not curated by RxNorm). This API function was released in June 2015 and is available at https://rxnav.nlm.nih.gov/.
Conclusion
The large percentage of recovered active concepts from obsolete NDCs in the three datasets is a positive indication that a service to identify obsolete NDCs and their active concepts from the past will be beneficial for the analysis of historical medication datasets. An smaller added benefit is the identification of active concepts for NDCs from other sources in RxNorm.

Acknowledgments
This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. We would like to thank Drs. Clem McDonald and Mallika Mundkur from the National Library of Medicine, who provided the Medicare use case. Similarly, our thanks go to Drs. Mark Homer and Ken Mandl from the Children's Hospital Informatics Program in Boston, who provided the use case with the private insurance.

References
Just One More Patient: Optimizing EMR Documentation in Ambulatory Care

Mark Pierce, MD, FAAP\(^1\), Tammy Toscos, MS, PhD\(^1,2\)
\(^1\)Parkview Health, Fort Wayne, IN; \(^2\)Indiana University School of Informatics and Computing, Indianapolis, IN

Abstract
The adoption of electronic medical records (EMRs) in primary care settings is on the rise in the United States and many are feeling the stress. The introduction of the EMR or transition to a new EMR is known to create workflow challenges for primary care providers and their office staff, as was the case in our health system. This study evolved out of an attempt to alleviate stress by defining the best practice or most optimal way to document office visits, allowing providers to see just one more patient each day. We leveraged a change management model that encourages looking for what is working vs. throwing resources at problem areas. By doing so we identified several distinguishing behaviors among providers who were doing exceptionally well with the EMR. We deployed an intervention aimed at enhancing the identified behaviors in a group of providers and it resulted marked improvement in efficiency.

Introduction
The transition to using an electronic medical record (EMR) in primary care offices can be quite difficult. In fact, several studies have shown that EMR adoption contributes to increased cognitive load, stress, burn out, and job dissatisfaction in primary care providers.\(^{1,2,3}\) In this paper we share the findings of a project aimed at identifying best practices for primary care office visits, as a means to improve the work experience of primary care providers working in our health system. We uncovered two key factors that differentiated providers’ use of the EMR for documentation. After implementing an intervention to enhance these two key factors, we found significant improvement in office visit documentation that occurred outside of business hours. Further, we found significant differences in specific EMR documentation actions of physicians who had the most successful office visits.

The Challenge
Parkview Health transitioned to a new EMR provider in the spring of 2012 and rolled out the application to all primary care offices in the system over the next two years. Going live with the EMR brought about much unhappiness. During the transition, providers were ‘slow and low.’ Providers complained of major delays in completing visit documentation and were becoming distressed about the impact on their office workflow and lengthy hours at work. Part of this was behavioral because there was not a strong pre-existing culture of EMR use, part was due to fatigue from the transition, and part was due to a reduction in productivity. Providers who had felt like an expert at all things medical before beginning use of EMR now struggled to get through basic work.

As Chief Medical Informatics Officer and a practicing primary care physician, the first author challenged the EMR company to provide a ‘best practice’ or most effective approach to completing all that needs to be done in an office visit. But when pressed for the best way the company could only respond by sharing what other organizations were doing – not an optimal method. In this particular EMR, there are many ways to complete a particular task, which allows for flexibility but also can create a cognitive burden and stress for the practicing physician.\(^4\) While the EMR training provided to primary care physicians was very task specific and technical, there was little attention paid to individual physician workflow. This left a gap in understanding what makes an office visit successful in the context of EMR use, what are the key metrics, and how do we get providers interacting with the EMR at an optimal level? The objective of our study was to enhance productivity, efficiency and quality of office visit documentation. In particular, the goal was to identify ways in which physician providers could spend less time in the system doing work and see just one more patient each day.

Interested in “What makes success?” vs. “What are the majority doing?”
It was a book on change management that inspired this work. In their book Switch, the Heath brothers describe a paradigm shift that can be affective when facing major obstacles against change, which they call finding the bright spots.\(^5\) In this analogy the authors describe a scenario where the problem solver looks in the community for successful efforts worth repeating instead of the common approach of deploying massive resources to those who
struggle. In the environment of change brought about by EMR adoption, a common approach is to focus all efforts towards low performers but this did not work for us. Things were not getting better and people who were doing well were being ignored. So we decided to look for the ‘bright spots’ in order to examine what was working well in the primary care practice use of the EMR. To do so, we identified several physician providers who seemed to be unfettered by the integration of the EMR in their practice. Importantly, we were not simply focused on what they did but rather on identifying what EMR practices were similar and which were true differentiators for those physicians practicing in the bright spots. In our data-driven approach we were careful not to use the words ‘high performers’ and ‘low performers,’ in order to avoid the obvious negative emotions that could be evoked. Rather, we relied on the metaphor ‘cheese finder’ used in another change management book, Johnson’s *Who Moved My Cheese* wherein finding cheese is synonymous with effectively navigating change.5

Methods
We conducted a 3-month study that included an observational pilot from which an intervention was designed. Outcome measures were collected in the final phase of the study to determine the impact of the intervention and to differentiate EMR actions of cheese finders when compared with controls.

Study Setting
Parkview Health is a system of 8 hospitals located in 6 counties in northeast Indiana. Parkview serves an area with a population of approximately 890,000. As a Health System, we have nearly 9,000 employees and 1,600,000 patient encounters each year. Parkview physicians group (PPG) has experience rapid growth and is currently comprised of 35 primary care service areas with 352 physicians (average age of 50) who are practicing across 120 locations. PPG serves counties in Northeast Indiana and Northwest Ohio with 1,253,000 clinic visits annually.

Observational Pilot
We started by identifying 12 cheese finders who met three criteria. 1) Productivity: Work Relative Value Unit (wRVU) ratio greater than 0.9; 2) Efficiency: less than two charts per week completed outside of regular office business hours; 3) Quality: content of note were complete and effectively communicated office visit events. We then matched 24 controls equally to the 12 cheese finders (2N:N) along specialty area and office setting (e.g. time since EMR deployed).

We developed a checklist of 35 items (see Figure 1) based discussion with cheese finders as to what they think makes them successful and also based on our observations of those doing well. Checklist information was obtained the information for all study subjects, either by direct observation in the office setting or by interviewing physician providers or their nurse. The objective of the checklist was to identify qualities that differentiated cheese finders from controls.

Intervention
The content of the intervention was focused on enhancing two actions that were identified in the observational pilot as a differentiating behavior of cheese finders. The intervention included two 5-minute educational videos deployed on Share Point. The videos were designed for everyone who was not a cheese finder as means to encourage behaviors that would lead to greater productivity. Participants were purely motivated by the potential to improve efficiency and recover personal time spent charting in the EMR. There were no other incentives to participate. Controls were individuals that either did not watch the videos describing the two interventions that could lead to higher productivity or they did watch the videos but chose not to change practice.

Outcome Measures
Outcome measures were selected keeping the end goal in mind: to enhance productivity, efficiency and quality of office visit documentation. In particular, the goal was to identify ways in which physician providers could spend less time in the system doing work and see just one more patient each day. Scope of EMR interaction was limited to documentation and ordering (to augment findings from observational pilot). Specific measures included the percentage of charts completed outside of office business hours and user interaction data collected with a tool offered by the EMR company – the user action log (UAL). The UAL tool allowed the capture of keystrokes, mouse clicks, menu interaction, and navigation sequences. So, this provided more detailed analysis of cheese finder behaviors than what was done in the observational pilot (e.g. time spent ordering, number of keystrokes per note). UAL data was collected for a 2-week period of time on both cheese finders and controls. Data analysis was completed using IBM SPSS Statistics for Windows, Version 21.0.
1. Is there any clinical "chart prep" completed? Example: Are charts reviewed prior to a scheduled appointment to look for recent test results or other missing data?
2. Does information get scanned prior to patient appointment or do you leave on paper for physician to review on paper for office visits?
3. Are there perceived bottlenecks at the front, impeding timely rooming of patient?
4. Does rooming staff fill in chief complaint?
5. Does rooming staff document vital signs every time?
6. Does rooming staff review and update allergies every time?
7. Does rooming staff do med rec?
8. Does rooming staff review smoking status?
9. Does rooming staff start the history of present illness (HPI)?
10. Does provider reference a snapshot view prior to entering exam room? Fill in blank
11. Are providers updating the problem list in the exam room?
12. Are smart sets being used?
13. For progress note documentation, is provider using a smart phrase for a whole note template?
14. Is provider using Dragon for HPI?
15. Is provider using Dragon for assessment/plan?
16. Is provider using Dragon for prescriptions?
17. Is provider using a system smart phrase to pull in lab results in the note?
18. Is provider using a personal smart phrase to pull in radiology results into the note?
19. Is provider using a prescription smart phrase?
20. Is provider using a review of systems (ROS) smart phrase?
21. Do you use a personal preference list for ordering?
22. Is provider using a macro for ROS?
23. Is provider using a macro for physical exam?
24. Is note being completed before seeing the next patient?
25. Does provider enter all of his/her own orders?
26. Is Dragon used for placing orders?
27. Are orders placed by provider in exam room?
28. Is visit diagnosis completed in exam room?
29. Is the problem list utilized for quick documentation of the diagnosis?
30. Is the level of service documented in exam room?
31. Is follow-up appointment documented prior to patient leaving?
32. Are notes routed by provider?
33. Is provider able to close the encounter at end of visit?
34. Does provider print own after visit summary?
35. Does clinical staff enter any orders in the system?

**Figure 1. Observational Checklist**

**Results**

The observational pilot findings uncovered two EMR actions that distinguished cheese finders from the rest of the cohort. A chi-square test was performed and a relationship was found between cheese finders and the frequency of using Dragon software to document the history of present illness, \( \chi^2(2, N = 32) = 5.722751, p = 0.016746 \). Likewise, a chi-square test revealed a relationship between cheese finders and the frequency of placing orders in the exam room, \( \chi^2(2, N = 32) = 4.097354, p = 0.04295 \). There were no other relationships found from the observational/ self-report data.

The intervention (two 5-minute educational videos) was deployed, focusing on encouraging the use of Dragon to document history of present illness and placing orders while in the exam room. Post-intervention measures showed a marked improvement of roughly 50% fewer charts being completed outside of office business hours (Figure 2).

![Figure 2](chart.png)

*Outside of time range 0800-1759 hours

**Figure 2. Percentage of Charts Completed Outside of Office Business Hours (OOBH) Pre- & Post- Intervention**

The findings from the UAL also revealed significant differences between the actions of cheese finders when compared to controls. Visit documentation metrics show cheese finders spent significantly less time \( p=0.028 \), fewer keystrokes \( p=0.002 \), and less clicks \( p=0.014 \) when documenting office visits (see Figures 3-5, respectively). Office ordering metrics show that while there was no significant difference in time spent ordering...
(p=0.126), cheese finders spent less time clicking (p=0.062) and had fewer clicks (p=0.061) in the process of placing orders when compared with controls (see Figures 6-8, respectively).

Figure 3. Difference between “Cheese Finders” and Control: Visit Documentation Time (p=0.028)

Figure 4. Difference between “Cheese Finders” and Control: Visit Documentation Keystrokes (p=0.002)

Figure 5. Difference between “Cheese Finders” and Control: Documentation Clicks/Visit (p=0.014)

Figure 6. Difference between “Cheese Finders” and Control: Ordering – Total Time (p=0.126)

Figure 7. Difference between “Cheese Finders” and Control: Ordering – Time Clicking (p=0.062)

Figure 8. Difference between “Cheese Finders” and Control: Ordering – Clicks/Visit (p=0.061)
Discussion
In this study we were able to demonstrate the value of applying change management philosophies during the aftermath of EMR implementation in primary care office settings. Namely, looking for the bright spots and uncovering exactly what efforts make an office visit successful and how we can communicate these practices to other physician providers who want to improve and/or simply reduce their cognitive load and stress from interacting with the EMR. Taking this alternative approach (the Switch model) to allocate resources toward the bright spots vs. toward those who were struggling allowed us to uncover two differentiating behaviors (using Dragon for HPI documentation and placing orders while still in the exam room) that were then leveraged in a successful intervention. The video education intervention resulted in a marked reduction in documentation completion outside of regular office hours. Additionally, we found that cheese finders spend less time documenting and completing notes with fewer keystroke and clicks. Although there was no real difference in provider time ordering, there was a tendency for fewer clicks among cheese finders. Drawing a connection between the findings from the observational pilot and the UAL findings we are left to wonder if using Dragon for documenting HPI can explain the time documenting and keystroke gap between cheese finders and controls.

There were confounding factors that impacted the findings of this study, namely organizational pressure to make things better right away. There was also a time barrier presented by an update in the EMR software and the looming deadline for the next go-live date. Additionally, there were contentions around allocating resources to examine behaviors of successful physician providers. There are limitations to the findings of this work in that our sample size was quite small. Further, we could not account for the time, clicks, and keystrokes entered by office RNs, LPNs and medical assistants.

We are currently looking at this opportunity to identify best practices for optimal EMR use with a larger, more formal study. In this we hope to define specific criteria to be considered a cheese finder, beyond the three criteria used for this study. This will include defining targets for documentation time, keystrokes for specific visit types, and ordering time. Expanded UAL data may include studying in-basket work, data gathering activity (e.g. navigation sequences) and behaviors when documenting care for inpatients. We also plan to expand our lens beyond the physician provider actions to better understand the role of support staff in the efficiencies we identified in cheese finders. We hope that we can identify more ideas for goal oriented training and optimization, as we found in this study. Finally, we hope to validate our findings with those from other organizations.

Conclusion
Are there times when “what everyone else does…” is not the right approach? We believe this is the case and have found value in applying activation resources to the bright spots in our own organization to determine what is right for us. Ultimately, this approach allowed us to quickly develop an effective intervention and allowed for more effective utilization of our resources.

Acknowledgements
This project would not have been possible without the contributions of Lindsey Swartz, Clint Keller, Tuan Newlin, Alexis Pierce, and Matt Cozon.

References
Casting a Wider Net: Data Driven Discovery of Proxies for Target Diagnoses

Dusan Ramljak, Adam Davey PhD, Alexey Uversky, Shoumik Roychoudhury, Zoran Obradovic PhD¹

¹Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA

Abstract

Background: The Hospital Readmissions Reduction Program (HRRP) introduced in October 2012 as part of the Affordable Care Act (ACA), ties hospital reimbursement rates to adjusted 30-day readmissions and mortality performance for a small set of target diagnoses. There is growing concern and emerging evidence that use of a small set of target diagnoses to establish reimbursement rates can lead to unstable results that are susceptible to manipulation (gaming) by hospitals.

Methods: We propose a novel approach to identifying co-occurring diagnoses and procedures that can themselves serve as a proxy indicator of the target diagnosis. The proposed approach constructs a Markov Blanket that allows a high level of performance, in terms of predictive accuracy and scalability, along with interpretability of obtained results. In order to scale to a large number of co-occurring diagnoses (features) and hospital discharge records (samples), our approach begins with Google’s PageRank algorithm and exploits the stability of obtained results to rank the contribution of each diagnosis/procedure in terms of presence in a Markov Blanket for outcome prediction.

Results: Presence of target diagnoses acute myocardial infarction (AMI), congestive heart failure (CHF), pneumonia (PN), and Sepsis in hospital discharge records for Medicare and Medicaid patients in California and New York state hospitals (2009-2011), were predicted using models trained on a subset of California state hospitals (2003-2008). Using repeated holdout evaluation, we used ~30,000,000 hospital discharge records and analyzed the stability of the proposed approach. Model performance was measured using the Area Under the ROC Curve (AUC) metric, and importance and contribution of single features to the final result. The results varied from AUC=0.68 (with SE<1e-4) for PN on cross validation datasets to AUC=0.94, with (SE<1e-7) for Sepsis on California hospitals (2009 – 2011), while the stability of features was consistently better with more training data for each target diagnosis. Prediction accuracy for considered target diagnoses approaches or exceeds accuracy estimates for discharge record data.

Conclusions: This paper presents a novel approach to identifying a small subset of relevant diagnoses and procedures that approximate the Markov Blanket for target diagnoses. Accuracy and interpretability of results demonstrate the potential of our approach.

Objective

Identify a small subset of diagnoses (Markov Blanket) that can serve as highly accurate proxies for the set of target diagnoses used to establish reimbursement rates under the Hospital Readmissions Reduction Program (HRRP). Identification of these subsets has applications to problems such as providing more stable hospital quality estimates, estimating the extent of diagnostic “gaming,” identification of potential “upcoding” or fraudulent claims, and fuller understanding networks of diseases and medical procedures.

Introduction

As of October 1, 2012, §3025 of the Affordable Care Act (ACA) dictates that hospital reimbursements have to be based on performance relative to preventable 30-day Medicare hospital readmission rates observed in hospitals with similar predicted risk profiles. Three specific diagnoses in particular are used to track reimbursement rates: acute myocardial infarction (AMI), congestive heart failure (CHF), and pneumonia (PN). In October 2014 chronic obstructive pulmonary disease (COPD) was added to the list of diagnoses with two additional procedures, but we did not include it in our experiments since its value in this context has been criticized²⁹. This policy change may have introduced an incentive for hospitals to under-diagnose these illnesses by substituting related diagnoses for which they will not be held accountable. Burgess and Hockenberry (2014)³¹ provide a historical perspective on current attention to hospital readmissions and consider the potential for gaming of readmissions, but not via subtle changes in diagnoses themselves. In addition to the three diagnoses we decided to analyze (AMI, CHF, and PN), we
included sepsis as a diagnosis for studying in our experiments, given that it is one of the most prevalent diagnoses, and is currently the diagnosis with the highest mortality rate in the US.

Our method is aimed at approximating Markov Blankets consisting of small subsets of diagnoses and procedures that frequently co-occur with, and can shield, each of our four target diagnoses from the rest of the disease/procedure network (in other words, the Markov Blankets for each target diagnosis node consist of diagnosis and procedure nodes that contain the only information required to accurately predict the behavior of the target diagnosis in question). We chose to use a Markov Blanket for this task since several studies have already shown that using only a small set of features that constitute the Markov Blanket for a dependent variable is sufficient to accurately predict the value of the variable. Hence, each target diagnosis can be accurately identified and inferred from a small subset of related diagnoses and procedures, which can then be used to identify true cases with target diagnoses, estimate the extent of gaming via substitute diagnoses/procedures, and also suggest related sets of diagnoses and procedures which, in combination, may provide more stable methods for setting reimbursement rates. We have observed that the most frequent co-occurrence by itself is necessary, but not sufficient, to establish an accurate approximation of a Markov Blanket.

We built a directed weighted network of diagnoses and procedures from the hospital records that contain our target diagnoses, and adjusted the weights in that network according to co-occurrences of diagnoses and procedures in the records that don't contain our target diagnoses. We apply Google’s PageRank algorithm and use the PageRank value as a criterion to identify important nodes which will belong to the Markov Blanket, and set the number of the “important” nodes to be selected based on the distribution of the number of diagnoses and procedures in the records.

We attained preliminary results that were accepted for presentation at SDM-DMMH 2015 Workshop which were tested on a very small subset of CA data. We have since then refined our approach to make it more scalable, and evaluated its validity on a much larger set of data, spanning two states (CA and NY) and several years. We found that our approach is both accurate and stable in both states, even when trained on a relatively small sample of CA dataset. Given its simplicity and interpretability, we are confident that it can be used effectively for any target diagnosis, not just the four that we focus on in our experiments.

**Background**

As a direct result of the change in the structure of Medicare reimbursements, there is now more focus on problems such as the ability of health care providers to identify changing predictors of 30-day hospital readmissions, as well as to identify characteristics of individuals and providers associated with above-average levels of readmission risk. Hospitals that perform below expectations will see a reduction of up to 1% in Medicare-based reimbursements for services related to all diagnostic-related groups (DRGs). Based on performance levels in 2010, these targets would have placed half of all hospitals in the under-performing group. In coming years, additional diagnoses will be added to the list used to determine reimbursement rates. A focus on 30-day readmission rates has been criticized for a variety of reasons, including concerns about the validity of diagnoses, sparse evidence that decreased readmissions translate into improved health outcomes, an assumption that most readmissions are preventable, and disproportionate penalization of hospitals serving “safety-net” hospitals. Joynt and Jha (2013) found evidence that large teaching hospitals and safety-net hospitals are more likely to be penalized under the HRRP. Kansagara et al. (2011) performed a systematic review of readmission risk models. They found that specific medical diagnoses were the most universally used predictors appearing in 24 of 26 prediction models they considered. This same study also found that the range of the Area Under the ROC Curve (hereinafter AUC) metric for predicting readmission ranged from 0.50 (nurse/case manager predicted risk of readmission) to 0.83 (administrative model plus self-report).

Several potential methods of “gaming” have been suggested, including enriching the population admitted with a target diagnosis with individuals assessed to have low readmission risk, using extended “holding areas” for patients to receive hospital care without being readmitted, and selectively coding target diagnoses among patients with expected low readmission risk. Such gaming strikes us as particularly likely given the narrow range of criteria that factor into reimbursement rates under the HRRP. Because hospitals cannot be penalized for diagnoses they do not make, physicians are incentivized to choose similar, but distinct, diagnoses for criterion diagnoses. For example, a patient who is admitted to a hospital with AMI may initially be diagnosed as having chest pains or coronary atherosclerosis. If this patient was subsequently readmitted within the following 30 days, this diagnosis could not be used to penalize the hospital for poor performance. Similarly, PN may initially be diagnosed as acute bronchitis or an upper respiratory infection, and CHF may instead be diagnosed at first as chronic obstructive pulmonary disease. In addition to studying the three aforementioned diagnoses used by the ACA, we also study Sepsis using our approach, since it is one of the most prevalent diagnoses, and is currently the diagnosis with the highest mortality.
rate in the US. It can be diagnosed initially as a bacterial infection, pneumonia, urinary tract infection, peritonitis, or a skin ulcer. In practice, only those assessed as having the lowest risk of 30-day readmission may be likely to receive the target diagnosis.

However, several specific diagnoses are likely to co-occur with the target diagnosis, and some procedures (e.g., angioplasty) may be strongly indicative of a specific underlying true diagnoses (e.g., AMI), serving as good proxy indicators of the true diagnosis. Evidence for changes to clinical practice, diagnoses, and associated procedures in response to changes in reimbursement has been well-documented for more than 30 years⁵ and there is reason to suspect that similar changes are already occurring due to the most recent changes enacted under the ACA. Rothberg et al. (2014)¹⁹ suggested that, by more liberally applying diagnoses of sepsis and respiratory failure, hospitals might improve their reported performance under the HRRP. Another way of estimating the extent of these changes, and identifying cases that represent the true diagnoses of criterion diagnoses, is considering diagnoses as a set of connected nodes in a graph (connected by aspects such as co-occurrence). The Markov Blanket (MB) for a node is the set of nodes that shield it from the rest of the network. Previous studies have shown that knowing the Markov Blanket of a diagnosis node is all that is required in order to predict the value of the criterion, either by classification or regression⁶,⁷,⁸. If the MB of a specific diagnosis can be identified prior to a policy change, it may be used to more accurately identify the set of criterion diagnoses following the policy change, which can in turn be used to estimate true cases, as well as the extent of gaming of diagnoses which will occur due to the policy change.

Our experiments used discharge data from the California and New York State Inpatient Databases (SID), obtained from the Healthcare Cost and Utilization Project (HCUP) provided by the Agency for Healthcare Research and Quality.⁶ The SID is a component of the HCUP, a partnership between federal and state governments and industry, which tracks all hospital admissions at the individual level. We included all data from January 2003 through December 2011. Patients were excluded from the analysis if they did not have Medicare or Medicaid as the primary payer and if they were younger than 19 years of age. The final dataset included 16,736,927 discharge records for CA and 12,717,787 discharge records for NY, with the primary set of features used in our experiences being the Clinical Classifications Software (CCS) diagnoses for ICD-9-CM. CCS codes, developed as part of the HCUP, are designed to cluster patient diagnoses (hereinafter DX) and procedures (hereinafter PX) into a manageable number of clinically meaningful categories (272 diagnoses and 231 procedure codes).

Some prior research has examined the role of comorbid conditions with the aim of identifying longer-term effects and mortality risk with a single target diagnosis in mind. Each of our target diagnoses has been considered in this fashion: AMI⁹, CHF¹¹, PN¹⁴, and sepsis¹⁵. To the best of our knowledge, ours is the first study concerned with identifying co-occurring diagnoses and procedures that can serve as a proxy indicator of the target diagnosis, something necessary to identify potential instances of hospitals gaming the system to reduce risk exposure.

Methods

Our goal was to find a minimum subset of the most informative DX and PX (accurately predict the presence of Target DX in the records) associated with the Target DX. We decided to build a directed weighted ego-centric network for each Target DX where weights are calculated from counts of co-occurrences of DX and PX. We calculated the PageRank for each DX, PX to obtain ranked list. We used the PageRank value as a criterion to rank the importance of DX, PX in the above defined Target DX ego-centric network. Number of features for PageRank approximation of MB identified as the maximum number of DX, PX in the records. This way obtained PageRank approximation of MB then serves as feature set for logistic regression with a default setup.²²,²³,²⁵. Generated software will be available at [https://github.com/dusanramljak](https://github.com/dusanramljak)

Target Diagnosis Ego-centric Network of Diagnoses and Procedures

We started with identifying DXs and PRs that most frequently co-occur with the Target DXs. The most frequent co-occurrence by itself is necessary, but not sufficient, to establish an appropriate approximation of the Markov Blanket. For example, some diagnoses might co-occur with a Target DX simply because they are frequently diagnosed. Meanwhile, other frequent diagnoses might co-occur with our Target DX and also contribute significantly in discriminating between the classes when viewed in combination with other diagnoses or procedures. Making that distinction is not possible by looking only at the frequencies of co-occurrences with a Target DX.

To that end we have built directed weighted networks of DX and PR from the hospital records that contained each of our Target DX. The starting weights in this network were counts of co-occurrences of the nodes. In order to build directed weights we followed the following intuition: the “level of trust” of a node (source) in a connection with another node (destination) should be scaled by the counts of occurrences of the source in the network – the source
can “trust” the destination only to an extent that is proportional to the ratio between the count of the occurrence of the link between them and its own count of occurrence. Since we also have the counts of occurrences of the nodes in the records that don’t contain our Target DX, we adjusted the weights in the network according to occurrences of destination in the records that don't contain our Target DX. Following this intuition, the “level of trust” should be scaled with the count of occurrences of the destination in the records that don’t contain our Target DX. The formula for setting the weights is $w = \frac{c_i}{c_s + c_d}$ where $c_i$ is the count of co-occurrences for the source and destination nodes, $c_s$ is the count of occurrences of the source node in the network, and $c_d$ is the count of occurrence of the destination node in the records that do not contain Target DX.

PageRank Approximation of Markov Blanket

Since we defined weights by co-occurrences, as well as by additional information from the structure of the network of DXs and PXs that co-occur with our Target DX, we could use the PageRank value as a criterion to identify important nodes. For a subset of highly important nodes, we could say that the nodes with highest PageRank represent an approximate Markov Blanket for our Target DXs.

PageRank gives us a ranked list of important DX and PX, but we are not able to determine if there are any redundancies. Redundancies in this context mean that several DX and PX might not provide information to discriminate between classes. In our earlier experiments we used a feature selection method to help us decide the number of DX and PX that will be provided by our PageRank approximation of MB, but our more recent experiments follow a different path. Because our goal is to have all the important DX and PX that could represent our Target DX in our MB, we determined the maximum number of DX and PX that could be present in individual records. That number is then used to determine the number of nodes with the highest PageRank included in the MB.

Results

The data we used in our experiments comes from the HCUP family of databases, and the raw data consists of patient hospital visit records from California’s and New York’s SID in the period from January 2003 up to December 2011. Each record consists of a number of attributes, which are explained in detail on the HCUP website\textsuperscript{26}. The California and New York database contain more than 50 million inpatient discharge records over the specified 9 years. The information is not specific to a group of hospitals, but rather represents the data for the entire state.

The database also includes demographic information for each patient (such as age, birth year, sex, race), DX (primary and up to 24 secondary for CA, primary and up to 14 secondary for NY), PX (up to 21 for CA, and up to 15 for NY), information about hospital stays, and other information (including length of stay, total charges, type of payment and payer, discharge month, and survival information). For the purposes of our experiments, we used only the procedure and diagnosis information since our earlier work showed its potential, and since one of our primary concerns is preserving privacy (which could be breached if we used the more specific demographic and hospital stay information). Each of the 4 Target DXs we examined in this study was fairly prevalent. As shown in Table 1, CHF was most common (17.37% CA, 16.51% NY), followed by PN (9.43% CA, 7.77% NY), Sepsis (6.13% CA, 4.91% NY), and AMI (3.25% CA, 3.14% NY). There were also considerable seasonal and secular trends (similar in both datasets) in these Target DXs. For example, the prevalence of sepsis increased steadily across the study period. The three other Target DXs showed gradual decreases in prevalence over time, but also very strong seasonal trends.

Table 1. Target DX frequency in CA and NY state databases

<table>
<thead>
<tr>
<th>Target DX</th>
<th>Sepsis</th>
<th>AMI</th>
<th>CHF</th>
<th>PN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>1,027,088</td>
<td>544,228</td>
<td>2,907,625</td>
<td>1,577,822</td>
</tr>
<tr>
<td>NY</td>
<td>625,310</td>
<td>399,371</td>
<td>2,100,602</td>
<td>988,475</td>
</tr>
</tbody>
</table>

We chose to use 20 combined DX and PX in this study, since this covered 95% of California and 98% of New York state data. We divided the data into two parts in order to show that the model generalizes well over both space and time. We used the first part, containing data from California in the years 2003 – 2008, for training and validating our models, and report the results of using 5-fold cross-validation. The second part consists of 2009 – 2011 records for both NY and CA data, and is used exclusively for testing. Since a previous study\textsuperscript{27} suggests that accuracy is
maximized when both training and test sets are balanced, used the same number of positive (Target DX present in the record) and negative (Target DX not present in the record) examples during training and testing. Since the four Target DXs appear at different rates (as shown in Table 1), the sample sizes we chose were different for each Target DX, but in each case we chose sizes with a roughly logarithmic progression such that the largest sample size covered the majority of the available cases. We used three sample sizes of different scale for each Target DX, and refer to these sample sizes using the general terms small, moderate, and large (for example, for CHF (2003-2008) in California, these sample sizes were 100k, 300k, and 1 million, respectively). We gathered 10 random samples for each sample size, and used 5-fold cross-validation for training and validation, resulting in 180 distinct training datasets per Target DX. When testing, we used the maximum possible size of the positive class for each Target DX, and performed 10 replications for each of the two states we have available, resulting in 20 distinct testing datasets for each Target DX. We opted to use 10 random samples that, when taken together, covered the entire dataset, rather than using the entire dataset as one big sample to increase the number of replications. To this end, we sampled cases with replacement with no overlaps within samples, and minimal to nonexistent overlap between samples.

For each of our Target DX, 150 out of 180 models were included in 5-fold cross-validation and tested once on an appropriate 5th fold holdout. 30 models were tested on all of the 20 test sets, and the results indicate high out-of-sample accuracy, scalability, and stability of the method. The fact that adding more features always increases the accuracy supports the idea that we are dealing with a challenging problem, though the diminishing returns on accuracy improvements suggest that using a smaller set of features can be sufficient. Furthermore, although diagnoses are the most commonly used features across different hospital readmission models, many other individual characteristics are also important but either not considered here (e.g., age, sex, race) or not available in the data we used (medications, laboratory tests ordered, laboratory test results). We include results using all diagnoses and procedures as features, rather than using only the significantly smaller subset of diagnoses and procedures selected by our method, to show that we attained comparable accuracy using only the subset of features that are strongly relevant to the Target DXs. Furthermore, experiments showed the somewhat surprising fact that having more data for training doesn’t offer much improvement in accuracy, but does affect the stability of both the achieved accuracy and the chosen features.

We evaluate the “stability” of the features selected by our method by examining how often each one appears in each of our training models. That is, after acquiring the ranked list of PageRank values for each of the features during each experiment, we selected the 20 highest-ranked features in each experiment (giving us 60 potentially overlapping sets of 20 features for each sample size), and then took the intersection of those feature sets to find features that appear in all experiments. We labeled these features as stable, since they were shown to be in the top 20 PageRank lists over the course of all experiments.

This procedure was repeated for each of the three sample sizes for each of the four target diagnoses, yielding the results summarized in Tables 2 and 3. These tables portray the features that were selected as the most influential by PageRank over each set of experiments, for each of the Target DX, as well as the sample size for which each feature was stable. In other words, features that are stable in all three sample sizes are the most frequently occurring and therefore are the best candidates for the MB approximation, while those that only appear during the experiments in the largest sample size are less frequent.

While accuracy is an important criterion for our task, interpretability and relevance of the selected features also play key roles in the utility and acceptability of the approach, and our results are very encouraging in this regard. For AMI (Table 2), the two selected diagnoses, cardiac arrest and shock, are two of the most common consequences of a heart attack. Similarly, the selected procedures all align directly with common clinical practice, representing a variety of cardiac imaging, diagnostic, and surgical procedures. CHF (Table 2) is a complex condition reflecting the intersection of cardiac and pulmonary systems. These are well represented among the selected diagnoses and procedures, as is the overlap (in terms of cardiovascular disease) with AMI. For PN (Table 3), one recent study found that readmissions rates for PN could be made more accurately and stably when sepsis and respiratory failure were also included. Both of these diagnoses are selected for PN, along with a range of respiratory procedures. Finally, the expected overlap between Sepsis (Table 3) and PN emerges for diagnoses and procedures when the former is considered as the target diagnosis. Additionally, there are more serious diagnoses relating to injuries and abscesses along with various stoma.

In addition to validating the relevance of the small subset of features we selected for each diagnosis, we also tested their representative power by comparing the predictive accuracy (measured in AUC) obtained using only this small subset of features against the accuracy obtained when using all (approximately 500) features for the same task.
These results are presented in Figure 3 as boxplots for each Target DX at each sample size, with variance obtained from the 10 repetitions of each experiment. Several observations can be made from these plots, the first being that while using all 500 features offers the best accuracy, we can attain very competitive predictive accuracy using a significantly smaller number (20) of features which have the added benefit of being relevant to the Target DX in addition to being useful for discriminating between the two classes. The gap between accuracies using the subset vs using all features varies among the four Target DX, but even in the case of the most difficult to predict diagnosis, PN, the AUC we attained using only 20 features was very respectable (in the 0.68-0.75 range). In the case of the easiest to predict diagnosis, Sepsis, we were able to attain a nearly perfect level of prediction using a very small number of features, which we believe to be particularly important given the frequency and deadliness of this diagnosis. A somewhat more surprising finding was that varying the sample size did not have as big of an impact as expected. Although a general trend of AUC being higher when the sample sizes are larger can be seen, there is a noticeable variability in the results, while the mean values of AUC are quite close between sample sizes. While we expected sample size to play a bigger role in increasing accuracy, these findings suggest that the features we chose were representationally powerful enough to offer good performance even when the sample size is limited, further proving their relevance to the target DXs.

Table 2. Table showing the diagnoses and procedures that were used in all experiments for appropriate sample sizes (marked by x in appropriate row) to form the PageRank approximated MB for AMI and CHF

<table>
<thead>
<tr>
<th>Size</th>
<th>Acute Myocardial Infarction</th>
<th>Size</th>
<th>Congestive Heart Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Diagnoses</td>
<td>I</td>
<td>Diagnoses</td>
</tr>
<tr>
<td>II</td>
<td>Cardiac arrest and ventricular fibrillation x x x</td>
<td>II</td>
<td>Acute and unspecified renal failure x x</td>
</tr>
<tr>
<td>III</td>
<td>Shock x x x</td>
<td>III</td>
<td>Acute myocardial infarction xx x</td>
</tr>
<tr>
<td></td>
<td>Procedures</td>
<td></td>
<td>Cardiac arrest and ventricular fibrillation xx x</td>
</tr>
<tr>
<td></td>
<td>Contrast aortogram x x x</td>
<td></td>
<td>Chronic kidney disease x</td>
</tr>
<tr>
<td></td>
<td>Contrast arteriogram of femoral and lower extremity arteries x</td>
<td></td>
<td>Conduction disorders xx x</td>
</tr>
<tr>
<td></td>
<td>Conversion of cardiac rhythm x x x</td>
<td></td>
<td>Heart valve disorders xx x</td>
</tr>
<tr>
<td></td>
<td>Coronary artery bypass graft (CABG) x x x</td>
<td></td>
<td>Hypertension with complications and secondary hypertension x x</td>
</tr>
<tr>
<td></td>
<td>Coronary thrombolysis x x x</td>
<td></td>
<td>Peri-; endo-; and myocarditis; cardiomyopathy xx x</td>
</tr>
<tr>
<td></td>
<td>Diagnostic cardiac catheterization; coronary arteriography x x x</td>
<td></td>
<td>Pulmonary heart disease xx x</td>
</tr>
<tr>
<td></td>
<td>Diagnostic ultrasound of heart (echocardiogram) x</td>
<td></td>
<td>Respiratory failure; insufficiency; arrest (adult) xx x</td>
</tr>
<tr>
<td></td>
<td>Extracorporeal circulation auxiliary to open heart procedures x x x</td>
<td></td>
<td>Shock xx x</td>
</tr>
<tr>
<td></td>
<td>Nuclear medicine imaging of pulmonary x</td>
<td></td>
<td>Procedures</td>
</tr>
<tr>
<td></td>
<td>Other non-OR therapeutic cardiovascular procedures x x x</td>
<td></td>
<td>Conversion of cardiac rhythm xx</td>
</tr>
<tr>
<td></td>
<td>Other OR heart procedures x x x</td>
<td></td>
<td>Diagnostic ultrasound of heart (echocardiogram) x</td>
</tr>
<tr>
<td></td>
<td>Other OR procedures on vessels other than head and neck x x x</td>
<td></td>
<td>Extracorporeal circulation auxiliary to open heart procedures x</td>
</tr>
<tr>
<td></td>
<td>Other therapeutic procedures x x</td>
<td></td>
<td>Heart valve procedures xx x</td>
</tr>
<tr>
<td></td>
<td>Percutaneous transluminal coronary angioplasty (PTCA) x x x</td>
<td></td>
<td>Insertion; revision; replacement; removal of cardiac pacemaker or cardio x</td>
</tr>
<tr>
<td></td>
<td>Respiratory intubation and mechanical ventilation x x x</td>
<td></td>
<td>Nuclear medicine imaging of pulmonary x</td>
</tr>
<tr>
<td></td>
<td>Swan-Ganz catheterization for monitoring x x x</td>
<td></td>
<td>Other OR heart procedures xx x</td>
</tr>
<tr>
<td></td>
<td>Tracheostomy; temporary and permanent x x x</td>
<td></td>
<td>Respiratory intubation and mechanical ventilation xx x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Swan-Ganz catheterization for monitoring xx x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tracheostomy; temporary and permanent x</td>
</tr>
</tbody>
</table>
Table 3. Table showing the diagnoses and procedures that were used in all experiments for appropriate sample sizes (marked by x in appropriate raw) to form the PageRank approximated MB for AMI and CHF

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>Size</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnoses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cystic fibrosis</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Other injuries and conditions due to external causes</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Pleurisy; pneumothorax; pulmonary collapse</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Respiratory failure; insufficiency; arrest</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Septicemia (except in labor)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Shock</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Procedures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arterial blood gases</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>CT scan chest</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Diagnostic bronchoscopy and biopsy of bronchus</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Enteral and parenteral nutrition</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Gastrostomy; temporary and permanent</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Incision of pleura; thoracentesis; chest drainage</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Other diagnostic procedures of respiratory tract and mediastinum</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Other non-OR therapeutic procedures on respiratory system</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Other respiratory therapy</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Respiratory intubation and mechanical ventilation</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Swan-Ganz catheterization for monitoring</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Tracheostomy; temporary and permanent</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Septicemia (except in labor)</th>
<th>Size</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnoses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspiration pneumonitis; food/vomitus</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Cardiac arrest and ventricular fibrillation</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Other injuries and conditions due to external causes</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Peritonitis and intestinal abscess</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Respiratory failure; insufficiency; arrest</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Shock</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Procedures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bone marrow transplant</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Colostomy; temporary and permanent</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Conversion of cardiac rhythm</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Enteral and parenteral nutrition</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Gastrostomy; temporary and permanent</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Ileostomy and other enterostomy</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Myringotomy</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Other non-OR therapeutic procedures on respiratory system</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Other non-OR therapeutic procedures on skin and breast</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Other vascular catheterization; not heart</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Respiratory intubation and mechanical ventilation</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Swan-Ganz catheterization for monitoring</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Tracheostomy; temporary and permanent</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

Conclusion

The US healthcare system is rife with opportunities for perverse incentives. The implementation of any new healthcare policy results in changes within the healthcare system in order to minimize the adverse consequences of the policy change for healthcare providers. Changes that began in 2012 under the Affordable Care Act can be expected to reduce the number of individuals receiving target diagnoses of AMI, CHF, and PN as healthcare providers move to reduce their exposure to adverse consequences of hospital readmissions. Beginning in October 2014, the HRRP was extended to three additional conditions: acute exacerbation of chronic obstructive pulmonary disease (COPD), patients admitted for elective total hip arthroplasty (THA), and total knee arthroplasty (TKA). Early models of readmissions are being presented for COPD30, but they suffer from the same low predictive value (AUC=0.65) as earlier work with AMI, CHF, and PN, and the value of COPD as a criterion measure has received some criticism.

In this paper, we propose a novel approach to the problem of under-diagnosing, specifically, approximating Markov Blankets by PageRank. Performance using this subset of diagnoses shows performance that is generally quite high in terms of both accuracy and precision. Additionally, these diagnoses and procedures often point to clinically meaningful patterns. However, it is unclear which will ultimately prove most useful as the network of diagnoses and procedures surrounding a Target DX changes in response to policy. To some extent, this problem is likely to pose a
continuously moving target and so future research should more fully develop an understanding of the temporal forces to determine whether, for example, the indicators of PN depend on month of admission.

**Figure 1.** AUC values of the model using the top 20 features selected by PageRank (blue boxes) and using all features (dotted red line). Each row represents one of the Target DX: AMI, CHF, PN and Sepsis. Each column represents a different subset of data that was used for testing purposes. Results obtained from the three sample sizes described above are shown for each setting.

The approach used here is likely to be useful in the analysis of healthcare data in several ways. First, it provides a set of associated diagnoses and procedures that can be used to “impute” missing or unobserved data in an effort to estimate true prevalence of various diseases. Second, it can be used to estimate the extent of “gaming” of diagnoses in response to policy changes. Several authors consider the potential for gaming. Multiple potential methods of “gaming” have been suggested, including enriching the population admitted with a target diagnosis with individuals...
assessed to have low readmission risk, using extended “holding areas” for patients to receive hospital care without being readmitted, and selectively coding target diagnoses among patients with expected low readmission risk. Such gaming strikes us as particularly likely given the narrow range of criteria that factor into reimbursement rates under the HRRP. Enriching the pool of related diagnoses has been shown to provide more accurate and stable estimates in the case of PN19.

Extending Rothberg et al. (2014)19, Sjoding et al. (2015)20 performed a Monte Carlo study using 2009 Medicare data. They found that hospitals could substantially improve their pneumonia readmission and mortality rates by converting pneumonia diagnoses to sepsis or respiratory failure. The improvements are often substantial. From a sample of 100 hospitals with pneumonia readmission rates above the 50th percentile, 66 improved their readmissions rate, and 15 dropped below the 50th percentile. Changes were even more dramatic when mortality was considered (90 and 41 hospitals, respectively). This suggests that our approach may also prove useful in order to adjust estimates for this kind of gaming and could also provide more robust methods to estimate true hospital readmission rates where intentional under-diagnosis of such diagnoses is likely. Historically, there are several precedents for this kind of under-reporting. The effects of the Omnibus Budget Reconciliation Act of 1987 (OBRA87) was observed to have considerable impact on medical practice in nursing home settings21. Considerable decreases in the per capita diagnosis of AMI can already be seen leading up to implementation of the HRRP (Gerhardt, et al., 2012).

Going forward, several areas of inquiry are likely to be fruitful. For example, future work should consider the conditions newly added in October 2014. Related to this, now that a reasonable amount of data are available following the implementation of the HRRP, researchers should evaluate evidence for shifting of diagnoses away from the targets. Another related area of critical importance has to do with the way in which individual hospital reimbursement rates are set and risk-adjusted. There is some evidence that building a larger network of related diagnoses may produce more stable and accurate performance estimates. Thus, a direct extension of this work is to consider the performance of a model trained on historical data, but used to predict performance of individual hospitals. Given how well our models generalize across time and state, we have reason to expect strong performance. Finally, models such as the ones we present here may have direct application to fraud detection and upcoding (Suresh et al., 2014)28.

Acknowledgements
This research was supported by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR and NSF through grant number NSF-14476570. HCUP, Agency for Healthcare Research and Quality, provided data used in this study.

References


Wei, Qiong ,Dunbrack, Jr, Roland L. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS ONE 2013; 8(3): e67863.


Impact of Robotic Surgery on Decision Making: Perspectives of Surgical Teams

Rebecca Randell, PhD1, Natasha Alvarado, PhD2, Stephanie Honey, PhD1, Joanne Greenhalgh, PhD1, Peter Gardner, PhD1, Arron Gill, BA3, David Jayne, MB BCh, MD, FRCS1, Alwyn Kotze, MB ChB, FRCA3, Alan Pearman, PhD1, Dawn Dowding, PhD, RN4+5
1University of Leeds, Leeds, UK; 2Newcastle University, Newcastle, UK; 3Leeds Teaching Hospitals NHS Trust, Leeds, UK; 4Columbia University School of Nursing, New York, USA; 5Center for Home Care Policy and Research, Visiting Nursing Service of New York, New York, USA

Abstract

There has been rapid growth in the purchase of surgical robots in both North America and Europe in recent years. Whilst this technology promises many benefits for patients, the introduction of such a complex interactive system into healthcare practice often results in unintended consequences that are difficult to predict. Decision making by surgeons during an operation is affected by variables including tactile perception, visual perception, motor skill, and instrument complexity, all of which are changed by robotic surgery, yet the impact of robotic surgery on decision making has not been previously studied. Drawing on the approach of realist evaluation, we conducted a multi-site interview study across nine hospitals, interviewing 44 operating room personnel with experience of robotic surgery to gather their perspectives on how robotic surgery impacts surgeon decision making. The findings reveal both potential benefits and challenges of robotic surgery for decision making.

Introduction

Technological innovation has led to great advances in surgical practice over the past two decades, resulting in improvements in patient outcomes1. In the 1990s, traditional open surgery was challenged by the introduction of minimally invasive surgery (MIS). With MIS, the surgeon performs operations using small ‘key-hole’ incisions, through which cameras and laparoscopic instruments are passed. This removes much of the abdominal access trauma, resulting in numerous benefits for patients, including less postoperative pain, shorter hospitalisation, quicker return to normal function, and improved cosmetic effect2-4. In addition to patient benefits, laparoscopic surgery is also cost-effective for healthcare providers5, due to shorter inpatient stay and decreased wound care costs4. However, laparoscopic surgery can be technically challenging to perform, as a result of the 2-dimensional operative image and instruments that have limited freedom of movement and require awkward and non-intuitive handling. As a consequence, uptake of laparoscopic surgery has been slow6.

Robotic surgery overcomes some of the limitations of laparoscopic surgery, potentially making the benefits of MIS available to a greater number of patients. The Da Vinci surgical robot (Intuitive Surgical, California, USA) is currently the only commercially available robot for soft tissue surgery. The surgeon sits unscrubbed at a console that provides them with a magnified pseudo 3-dimensional (3D) view of the surgical site. From the console, the surgeon is able to control the robot arms that hold the laparoscopic instruments inserted into the patient. Robotic surgery enables the surgeon to achieve increased precision through intuitive instrument handling, tremor elimination, and motion scaling. There has been rapid growth in the purchase of Da Vinci robots in North America, despite the cost of the latest model being almost $2 million and annual maintenance costs of $125,000, and Europe is quickly following suit. Between 2007 and 2011, the number of Da Vinci robots installed in the United States increased from 800 to 1,4007, while the number of Da Vinci robots installed worldwide had reached 2300 in 20118. Whilst robotic surgery is primarily used in urology, its use is expanding across the surgical specialties, also being used in gynecology, ear nose and throat, colorectal, cardiology, and pediatrics.

Decision making is an important component of surgical expertise9, yet there is a paucity of research on decision making in the operating room (OR)10, 11. Theories of decision making highlight the importance of previous experience with particular situations, to enable the development of patterns or mental models which draw attention to relevant cues, provide expectations, determine plausible goals, and suggest typical responses to the situation12. In the OR, factors that impact surgeons’ decision making include tactile perception, visual perception, motor skill, and instrument complexity13, all of which are affected by robotic surgery, and may therefore impact on the surgeons’ ability to use their experience (mental models) to inform decision making. Similarly, the separation of the surgeon
from the rest of the OR personnel may also be significant, as there is strong evidence, both in the OR and in other contexts, that physical proximity of team members and technology influence the gathering of information that is used to inform decision making. Again, disruption of these patterns of information seeking may impact if and how surgeons are able to use their experience to inform their decision making. However, evaluation of robotic surgery to date has understandably focused on patient outcomes and the impact of robotic surgery on decision making has not previously been studied.

This is the first study to explore how robotic surgery impacts decision making, through interviews with surgeons and OR personnel. Interviews were conducted using the teacher-learner cycle, where interviewees are presented with ideas from the literature and asked to reflect on the extent to which those ideas fit with their experience. We first review relevant literature on decision making before describing the methods of our study and presenting the results. We conclude by discussing the implications of the findings for the implementation of robotic surgery.

Background

Klein, in his recognition primed decision (RPD) model, highlights the importance of context or situation in ‘triggering’ mental models that guide decision making in numerous complex decision situations. Situation awareness is defined as the perception of elements in the environment, the comprehension of their meaning, and the projection of their status in the near future. Situation awareness is also an important factor in understanding information behavior, the manner in which decision makers seek and use information to guide their choices. As highlighted in the RPD model, situation awareness is an important component of surgeons’ intra-operative decision making and better situation awareness of the surgeon is associated with fewer surgical errors. One model of intra-operative decision making suggests a continuous cycle where, with the preoperative plan in mind, the surgeon assesses the situation, reconciles new information with existing information, and subsequently implements a revised course of action. In this cycle, through the use of existing mental models, information may be actively sought or, by remaining observant of what is happening in the OR, perceived without active seeking. Robotic surgery potentially changes the nature of the information that the surgeon has available to them. The magnified, surgeon-controlled, 3D view of the surgical site may support the surgeon in visually perceiving anatomic information. However, because the surgeon sits at a console away from the patient and the rest of the OR team, they may not be able to see the patient or the robot arms directly and access to auditory information is also likely to be reduced. One report of the introduction of robotic surgery described a tendency for surgeons to ‘bury themselves in the console,’ thereby blocking out the OR. Consequently, the surgeon is dependent on the rest of the OR team communicating information that they previously obtained through visual perception. This has led some to argue that intra-operative decision making in robotic surgery is more collaborative than open or laparoscopic surgery.

Robotic surgery also changes the ability of the surgeon to use tactile perception to determine anatomic information. In open surgery, surgeons work primarily with visual and tactile information. In laparoscopic surgery, tactile information is reduced but, by touching with the instruments, surgeons are still able to determine features of objects such as shape, texture, and consistency. In robotic surgery, the surgeon receives no tactile information and this is considered to be a major limitation of robotic surgery. Some surgeons have suggested that the lack of tactile information means that surgeons move more slowly because they have to rely on visual information only. However, research suggests that, as experience of robotic surgery increases, surgeons find visual information sufficient for informing their intra-operative decision making, and this is supported by surgeons’ own reports.

While robotic surgery changes the visual and tactile information that is available to the surgeon, potentially reducing their situation awareness, it has been argued that the surgeon’s position inside the console and the 3D image create a sense of immersion and that the subsequent reduced ‘distractibility’ of the surgeon could be a benefit of robotic surgery. Certainly, if robotic surgery does reduce the number of distractions that the surgeon experiences, this could have positive impacts on decision making and subsequent patient outcomes; research reveals that distractions in the form of case-irrelevant communication are linked to an increase in surgeons’ mental fatigue and intra-operative stress and excessive levels of intra-operative stress compromise not only technical skills but also non-technical skills such as decision making. However, it has also been found that equipment and work environment distractions are more frequent in laparoscopic operations than in open operations, due to the more complex technology, suggesting that robotic surgery may also introduce new distractions.

The ergonomic benefits of robotic surgery may also impact surgeons’ stress and fatigue. Robotic surgery removes the awkward and unnatural movements required during laparoscopy and the surgeon is able to sit down comfortably at the console, potentially reducing physical stress and associated fatigue. This has led some surgeons to argue that, with stress arising from a difficult operation being an indirect cause of conversion to open surgery, robotic surgery
may result in a lower rate of conversion. However, results from experimental studies are inconclusive; while two studies found robotic surgery to result in lower mental and physical stress than laparoscopic surgery, in another study the difference was not statistically significant.

Methods

This study was undertaken as part of a process evaluation, running alongside ROLARR, a multicenter randomized controlled trial comparing laparoscopic and robotic surgery for the curative treatment of rectal cancer. Realist evaluation, an approach that is increasingly popular for the evaluation of complex interventions in healthcare, provides an overall framework for the process evaluation. Realist evaluation involves building, refining, and testing users’ ideas and assumptions, or ‘theories’, of how an intervention produces its outcomes. From a realist perspective, interventions in and of themselves do not produce outcomes. Rather, interventions offer resources to users; outcomes depend on how users choose to respond to those resources, which will vary according to the situation or ‘context’. Thus, rather than just asking ‘what works?’, realist evaluation seeks to answer the question of ‘what works, for whom, in what circumstances, and how?’. Realist evaluation involves gathering data in order to explain how different contexts trigger particular changes in the reasoning and responses of users (‘mechanisms’) which, in turn, give rise to a particular pattern of outcomes. While general qualitative approaches can only provide a catalogue of possible contextual factors thought to impact the process and outcomes of interest, the advantage of realist evaluation is that it increases the specificity of our understanding of the relationship between context, mechanisms and outcomes. As part of the first phase of the process evaluation, we undertook a multi-site interview study, interviewing OR teams to elicit their theories of how robotic surgery impacts decision making during surgical operations.

Participants

All English hospitals participating in the ROLARR trial were invited to participate in the interview study. English hospitals not participating in the trial but using the robot for colorectal surgery were identified by the trial team and through personal contacts of one of the team members (DJ), and all were invited to participate in the interview study. In this way, geographic spread and variation in level of experience of robotic surgery was achieved. A snowball sampling strategy was used; at each hospital, one of the surgeons was interviewed first, who then assisted in identifying other members of the OR personnel to interview (surgeons, trainee surgeons, anesthesiologists, OR nurses, and OR practitioners).

National Health Service (NHS) study-wide ethical approval was granted and research governance permissions were obtained from each hospital. All participants gave informed consent.

Data collection

Interviews were conducted, either face-to-face or by telephone, using the teacher-learner cycle. Teacher-learner cycle interviews are advocated within realist evaluation as a way to uncover users’ ideas and assumptions (theories) about how an intervention works and thus understand how user responses (mechanisms) are triggered in different circumstances (contexts) and produce certain outcomes. These theories can then be expressed as context-mechanism-outcome configurations (CMOs). In contrast to standard qualitative interviews, in teacher-learner cycle interviews, the researcher’s theories are the subject matter and the purpose of the interview is to confirm, falsify, or refine that theory. Using a semi-structured interview schedule, the interviewer presented the interviewee with theories from the literature concerning how robotic surgery is thought to impact surgeons’ decision making. Interviewees were asked to reflect on whether or not, and in which ways, these theories fitted with their experience.

From our review of the literature, we started with the following theories:

1. When the team is more experienced in robotic surgery, they understand that the surgeon’s situation awareness is dependent on them orally communicating information and they respond by using more oral communication about the patient’s state which in turn improves the surgeon’s situation awareness.
2. Surgeons progress more slowly through a robotic procedure because they do not have tactile information to inform their assessment of the situation and to determine whether to persist with or revise their course of action, but this effect becomes less pronounced as experience with robotic surgery increases.
3. The sense of immersion that the robot provides means that the surgeon is more focused, resulting in improved decision making and patient outcomes.
4. The ergonomics of the robot mean that the surgeon is less stressed and tired, resulting in better decision making and reduced conversion to open surgery.
As the literature provided limited information on the contexts in which robotic surgery impacts decision making, and the mechanisms through which those impacts are achieved, the interviews sought to elicit further detail about this. An iterative approach to data collection and analysis was taken, with the theories being revised as the interviews progressed. Interviews were audio recorded and transcribed verbatim.

**Analysis**
The interview transcripts were anonymized and entered into NVivo 10. Framework analysis, an approach developed for analyzing qualitative data for applied policy research, was used. Informed by the interview schedule and reading of preliminary interviews, codes for indexing the data were identified and agreed by three members of the research team (RR, NA, SH). They then indexed four transcripts to test the applicability of the codes and assess agreement. Where there was variation in the indexing, the codes were refined and definitions were clarified. The refined codes were applied to all transcripts. The indexed data was summarized in a matrix display to build up a picture of the data as a whole. In the final stage, mapping and interpretation, the matrix was used to identify similarities and differences in participants’ responses.

**Results**
Forty-four interviews were conducted across nine hospitals between January and August 2014. Interviews ranged from 29 minutes to 1 hour 40 minutes, with an average (mean) length of interview of 53 minutes. Table 1 provides a summary of participants and settings. The findings are organized around the main theories discussed.

**Table 1.** Participants by professional group and hospital type.

<table>
<thead>
<tr>
<th>N = 44</th>
<th>N (% of sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Professional group</strong></td>
<td></td>
</tr>
<tr>
<td>Surgeon</td>
<td>12 (27)</td>
</tr>
<tr>
<td>Trainee surgeon</td>
<td>5 (11)</td>
</tr>
<tr>
<td>Manager</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Anesthesiologist</td>
<td>6 (14)</td>
</tr>
<tr>
<td>OR Nurse</td>
<td>13 (30)</td>
</tr>
<tr>
<td>OR Practitioner</td>
<td>7 (16)</td>
</tr>
<tr>
<td><strong>Hospital type</strong></td>
<td></td>
</tr>
<tr>
<td>Teaching</td>
<td>21 (48)</td>
</tr>
<tr>
<td>District general</td>
<td>17 (39)</td>
</tr>
<tr>
<td>Cancer center</td>
<td>6 (13)</td>
</tr>
</tbody>
</table>

**Theory 1: Situation awareness**
The majority of surgeons perceived that their situation awareness is potentially reduced during robotic surgery, stating that they are focused on a small area and therefore are less aware of their environment; they have ‘tunnel vision’. One surgeon provided the example of ‘sucking fluid’; the surgeon can request that their assistant provide suction, but they are unaware if the assistant experiences difficulties fulfilling their request or the reasons why. Attitudes to the seriousness of this varied. One surgeon described operating with a second surgeon as a ‘wing man’ to counteract the problem of reduced situation awareness and stated that he would be very concerned about reduced situation awareness if he operated without a second surgeon present. Another surgeon described being ‘vastly less aware’ of what is going on in the OR but he had not been ‘hindered’ by this and did not think it made any difference to the operation. Only two surgeons felt that their situation awareness is not reduced. One described still listening to the ‘banter’ amongst the team in the OR, while the other surgeon made a conscious effort to intermittently ask the team about the patient’s status.

Team members also perceived that the surgeon’s situation awareness is reduced due to their position in the console e.g. the surgeons do not have lateral vision and their sensory feedback, which can indicate problems, is reduced. Consequences of the surgeon’s reduced situation awareness described by the team include the robotic arms impinging upon each other, which could damage the robot or prevent the surgeon achieving their aim. Respondents explained that the surgeon only realizes that the robot arms are clashing when they are unable to manipulate the instruments as desired. On one occasion the robot arms nearly collided with a patient’s head; this problem was averted by the OR nurse who intervened.
The overarching strategy described by the surgeons to increase situation awareness, in line with our initial theory, was to establish good communication links between the surgeon and the team. Good communication was seen as an essential part of robotic surgery. Trust between the surgeon and the team was also emphasized, as the surgeon has to rely on the rest of the team to communicate information outside of their field of vision to avoid complications. If the surgeon trusts their team to communicate problems to them, their concern over their reduced situation awareness is lessened. One surgeon commented that a more experienced team might be better able to communicate the necessary information to them.

Communication was also described by the OR teams as the main strategy to increase situation awareness, who saw it as their responsibility to act as the ‘surgeon’s eyes and ears’. In contrast to our initial theory, the information that they described communicating to the surgeon was less about the patient state and more often about the robot, as in the examples described above. Some noted that they just ‘tell the surgeon’ when there are problems and that everyone in their team knows to do this whether it is themselves or the robot that is struggling. Others described that it is important, because of the physical separation of the surgeon from the team, that team members have voices that are ‘strong enough for the surgeon to hear’. Good communication was seen as dependent on the relationship between the surgeon and team. While this is dependent on individual personalities and approaches, training together as a team and having a dedicated team were strategies that were considered by interviewees to increase team members’ confidence to speak up. As one nurse said about training as a team:

‘I just think having been away, just you got to know the surgeons better and […] you’re just that much happier saying, can we start, can we slow down a bit, can we do this or we need to get that, it just sort of levels the hierarchy so much, which made it much easier to work with people.’

Another strategy described by the OR teams was positioning the console so that the surgeon has a direct view of the patient and the assistant when they look up from the robot, i.e. they are ‘not hidden in a corner.’

Theory 2: Lack of tactile information

Several surgeons described initial experiences with the robot where, due to the absence of tactile information, they had not realized how much force they were applying and consequently had, for example, snapped a suture. However, none of the surgeons considered the lack of tactile information a significant problem. While a couple of the surgeons described being ‘a bit more careful’, ‘a bit more hesitant’, the surgeons we interviewed did not consider that the lack of tactile information led to a longer operation duration. They felt that they had adapted quickly to relying on visual cues, learning to look for tension. Several surgeons related this to their experience of laparoscopic surgery; with laparoscopic surgery, they had already learnt to work with reduced tactile information. As one surgeon described, ‘from previous experience you know what you’re looking for, so you know the tension that you’re putting on the tissues from what you can actually see.’ Interviewees contrasted this with the experience of urology surgeons who had moved straight from doing open prostatectomies to doing them robotically.

Theory 3: Immersion

The majority of the surgeons we interviewed agreed that the robot produces a sense of immersion. One surgeon described how they can ‘lose themselves’ during the operation and, referring to level of concentration, he described this feeling as ‘quite intense’. Other surgeons commented that it is not that the robot creates a sense of immersion but just that they have to concentrate more because they have less experience with robotic surgery than with laparoscopic surgery. Two surgeons refuted the idea that the robot produces a sense of immersion, commenting that they are immersed in the procedure regardless of whether it is laparoscopic or robotic and that technology should not determine whether the surgeon is immersed.

A number of theories about the contexts in which a sense of immersion occurs were suggested by the participants. One surgeon anticipated that, while he already experiences a sense of immersion when using the robot, the feeling will probably increase when the ‘mundane’ and routine tasks related to using the robot, e.g. port positioning, have been mastered. In contrast, another surgeon commented that he feels immersed using the console, particularly during complex cases, but that this feeling would probably lessen over time, i.e. that it was a feature of his limited experience with the robot. One surgeon described immersion as being dependent on who he has assisting i.e. if he trusts the assistant he can be immersed as the assistant fulfils requests with ‘silver service’, whereas otherwise he is ‘constantly looking’ as there is anxiety about where the assistant is ‘pointing the instrument’. The creation of trust was also associated with training as a team, as one surgeon described:
'We learned to trust each other. We came back from [the training] with that certain knowledge that between us we knew what we knew and [...] we would each remember something and we would be able to pull it off.'

Some surgeons described the OR as quiet during robotic surgery, enhancing their concentration, and that there are no distractions. In comparison, in open and laparoscopic cases, the surgeon can chat with the assistant and team.

Perceptions of the impact of the sense of immersion varied. Some respondents commented that heightened concentration might lead to better decision making, but how or why was not articulated. One surgeon described the sense of immersion as making him more focused, which should enable a more precise dissection. Others felt unable to comment on whether immersion would be reflected in patient outcomes. One surgeon said he felt the sense of immersion would not impact his decision making, except that he may persevere longer with an operation because he is less aware of time. However, this could cause concern if the patient is operated on for an ‘excess amount of time’.

**Theory 4: Impact of ergonomics**

The surgeons discussed their experiences of using the robot and the extent to which ergonomics affected their levels of stress and tiredness in comparison to laparoscopic surgery. Some surgeons discussed that, for them, performing operations using the robot is more stressful than laparoscopic surgery because they are in the early stages of implementation i.e. have not used the robot on many occasions. In this context the surgeons stated that they shared the operation with a colleague. They explained that sharing the operation reduced their levels of stress, as opposed to the ergonomics of the robot. Other surgeons felt that the robot was an improvement on laparoscopic surgery (ergonomically); how and why it was an improvement was not fully explored although one surgeon described that they were in a ‘less awkward position’. The surgeons also discussed that using the robot might be physically less tiring than laparoscopic surgery, but it is mentally more so because they have less experience of robotic surgery than laparoscopic surgery. For this reason, they have a higher level of concentration for a longer time using the robot; in comparison, they could relax on occasion during laparoscopic cases.

Two surgeons described how the level of stress is affected by the how the team acts; as one surgeon described it:

‘I think it probably makes you physically less tired. I think you’re probably mentally more tired [...] We’ve all done less robotics than we have laparoscopic, so you’re carrying more of a burden I think robotically. And sometimes you feel like you’re only the person in theatre that knows what’s going on. [...] Because you’re there, and you’re the only one there looking.’

The extent to which the robot reduces surgeons’ stress levels was also described as dependent on the stage of the procedure. For example, talking about suturing, one surgeon said:

‘If I was doing that laparoscopically, it would be a nightmare. It’s just a joy to do it robotically because of the ergonomics.’

In contrast, this surgeon described how with dissection his ‘fear’ of bleeding is increased, because the magnified image means that he notices tiny blood vessels that he would not notice otherwise. It was also noted that stress can be dependent on the type of operation performed e.g. a low anterior resection is stressful using both approaches, whereas operations that do not go down to the pelvic floor are less stressful and demanding. However, being able to take breaks when using the robot was noted as a benefit.

The extent to which the ergonomics of the robot impacted on decision making, particularly the decision to convert to open surgery, was difficult to ascertain. It was suggested by some respondents that if the surgeon is more comfortable during surgery he or she might persevere with a difficult operation rather than convert to open surgery. It was also noted that the surgeon can ‘take five minutes’ to consider their decisions during robotic surgery, whereas they might feel more pressure in decision making during laparoscopic surgery. However, it was also acknowledged that the decision to convert to open surgery is often due to circumstances outside the surgeon’s control, e.g. conversion was described as a ‘technical’ matter that was not linked to ergonomics or how stressed the surgeon was. One surgeon stated that, if anything, he would persevere longer with laparoscopic surgery because that is the technique with which he has more experience and so feels more confident.

How and why the ergonomics of the robot reduced surgeons’ stress levels was also postulated by the wider surgical team. Team members discussed a number of ergonomic benefits of the robot, e.g. because the surgeon is sat down must mean that they are more relaxed, the surgeon can adjust the console’s head piece, the console is padded, and it is easier to have coffee breaks as no scrubbing or de-scrubbing is required to step away from the console. However, it was also suggested that stress might be dependent on the surgeon’s experience i.e. those learning how to use the
robot do not seem as relaxed using the console. Participants also noted that if a surgeon found a stage of the procedure difficult, this would cause stress regardless of the ergonomics of the robot. The difference between mental and physical tiredness was also highlighted by the team; some described that the surgeon gets tired looking at the 3D image, that robotic surgery is stressful for their eyes and requires more mental concentration. It was also noted that the surgeon can be hunched in the same position for hours.

Discussion

Robotic surgery is a complex interactive system. Whilst this technology promises many benefits for patients, the introduction of interactive systems into healthcare practice often results in unintended consequences that are difficult to predict. We have drawn on the experience of OR teams to understand the impact that robotic surgery has on surgeons’ intra-operative decision making. Using the approach of realist evaluation, we have not only identified some of the consequences of robotic surgery for the processes and outcomes of surgeon decision making but have also begun to unpack how these impacts are achieved and the contexts in which these impacts are likely to occur.

The findings suggest a number of revisions to the theories discussed in the interviews. They highlight the role of the team in maintaining the surgeon’s situation awareness, fitting with ideas previously postulated in the literature. We anticipated this would involve teams providing surgeons with information about the patient state but the state of the robot also needs to be communicated. However, the findings also suggest that, for this to occur, there needs to be a positive relationship between surgeon and team. That relationship may be impacted by the way in which robotic surgery is introduced, with training as a team and having a dedicated robotic team being associated with positive relationships between the surgeon and team so that team members feel confident to speak up. The findings also reveal the intertwined nature of surgeon situation awareness and the surgeon’s level of concentration when undertaking robotic surgery; when the surgeon trusts the team to make him aware of changes outside of his field of view, he feels confident to remain in the console, resulting in reduced distraction and increased concentration. What is less clear is how this impacts patient outcomes. The ergonomic console can reduce stress and tiredness, enabling the surgeon to persist longer with the operation and potentially reducing the number of conversions to open surgery, but this is only when the surgeon is experienced in robotic surgery. Interestingly, in contrast to some of the literature, lack of tactile information did not present a concern for the surgeons in our study. A revised set of theories, formulated as CMO configurations, is presented in Table 2. Given realist evaluation’s concern with identifying what works, for whom, in what circumstances, these theories describe what is needed to produce a positive outcome. The implication is that, in the absence of the necessary contextual factors, the mechanism that produces the desire outcome will not be triggered. For example, if there is not a positive relationship between the surgeon and the OR team so that the team communicate information to the surgeon, this could lead to complications and increased distraction for the surgeon.

Table 2. Revised theories presented as CMO configurations.

<table>
<thead>
<tr>
<th>CONTEXT</th>
<th>RESOURCE</th>
<th>MECHANISM</th>
<th>RESPONSE</th>
<th>OUTCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive relationship between surgeon and OR team</td>
<td>Team communicates to surgeon information about patient and robot</td>
<td>Surgeon adjusts their course of action based on the information</td>
<td>Surgeon feels confident to remain in console</td>
<td>Reduced distraction and increased concentration</td>
</tr>
<tr>
<td>Surgeon experienced in robotic surgery</td>
<td>Ergonomic console</td>
<td>Surgeon feels comfortable to persist longer with the operation</td>
<td>Reduced levels of stress and tiredness</td>
<td>Reduced conversion to open surgery</td>
</tr>
</tbody>
</table>

These findings have a number of implications for the design and implementation of surgical robots. While there is recent research exploring how to provide haptic feedback in robotic surgery, from the perspective of the users the lack of tactile information does not, after a short learning period, hinder their ability to assess the situation and...
determine the appropriate course of action. Concerns that are more persistent relate to the impact on the surgeon’s situation awareness and, where this is not addressed, potential benefits of robotic surgery in terms of reduced distraction and increased concentration will not be obtained. We suggest that, to realize the benefits of robotic surgery for surgeon decision making and avoid any negative consequences, implementation of robotic surgery should involve (a) training for teams that acknowledges the need for the team to maintain the surgeon’s situation awareness, and (b) whole team training and/or a dedicated robotic team to establish positive strategies of communication between the surgeon and the team.

Limitations

A limitation of this research is that, although conducted over nine different hospitals, it has been concerned with one surgical specialty, colorectal surgery. However, informal discussions with urology and gynecology surgeons suggest that they experience similar impacts of robotic surgery. In future research, we will be conducting interviews across a range of surgical specialties to assess the extent to which are findings are specific to colorectal surgery and to revise our CMO configurations to reflect the experience of a broader range of surgical specialties.

Another limitation of this research relates to the challenges associated with conducting interviews to understand decision making. Decision making in the OR has predominantly been studied through interviews but to develop a rich, nuanced understand of the complexity of clinical decision making requires comprehensive data, gathered through multiple methods. Consequently, in the next phase of this research, we will be testing our revised theories through a multi-method multi-site case study. We will be conducting structured observations of both robotic and laparoscopic operations using OTAS (Observational Teamwork Assessment for Surgery), which will provide a quantitative measure of the situation awareness of the different sub-teams in the OR. Post-operation, we will ask participants to complete questionnaires to gather their perceptions of the mental and physical demand and the extent of distractions. This will be complemented by detailed analysis of video data that allows us to understand how these impacts are achieved, interviews with participants to understand their reasoning, and ethnographic observations to understand the contexts that influences these mechanisms.

Conclusion

This is the first study to explore how robotic surgery impacts decision making. It reveals both potential benefits and challenges of robotic surgery for decision making, which could have consequences for patient outcomes. While the assumption underlying the introduction of robotic surgery is that the increased precision provided by the robot results in improved patient outcomes, our findings suggest a more complex picture. This is a topic that needs to be considered and addressed by healthcare providers when implementing robotic surgery into their organization.

Acknowledgements

We would like to thank the surgeons and OR personnel who generously gave up their time to be interviewed. This research is funded by the National Institute for Health Research (NIHR) Health Services and Delivery Research (HS&DR) Programme (project number 12/5005/04). The ROLARR trial is funded by the Efficacy and Mechanism Evaluation (EME) programme, which is funded by the Medical Research Council (MRC) and managed by the NIHR. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the HS&DR Programme, NIHR, MRC, NHS or the Department of Health.

References

AMIA members’ “vital signs”: what the HIT implementation listserv says about goals for AMIA and for medical informatics

Kourosh Ravvaz, MD, PhD, MPH\textsuperscript{a}, Craig Kuziemsky, PhD\textsuperscript{b}, Ross Koppel, PhD, FACMI\textsuperscript{c}, Bonnie Kaplan, PhD, FACMI\textsuperscript{d}, Samantha A. Adams, PhD\textsuperscript{e}, Martha B. Adams, MD, MA, FACP\textsuperscript{f}*

\textsuperscript{a}Aurora Research Institute, Aurora Health Care, Milwaukee; College of Health Sciences, University of Wisconsin, Milwaukee, WI, USA
\textsuperscript{b}Telfer School of Management, University of Ottawa, Ottawa, ON, Canada
\textsuperscript{c}Sociology Department; Center for Clinical Epidemiology and Biostatistics, Medical School; LDI, Wharton School, University of Pennsylvania, Philadelphia, PA, USA
\textsuperscript{d}Yale Center for Medical Informatics at Yale University School of Medicine; Yale Interdisciplinary Center for Bioethics; The Information Society Project at Yale Law School, Yale University, USA
\textsuperscript{e}Tilburg Institute for Law, Technology, and Society, Tilburg Law School, Tilburg University, The Netherlands
\textsuperscript{f}Division of General Internal Medicine, Department of Medicine; Division of Clinical Informatics, Department of Community and Family Medicine; Duke Center for Health Informatics, Duke University Medical Center, USA

*authorship in reverse alphabetical order

Abstract

The health information technology (HIT) implementation listserv was conceived as a way to combine a substantial portion of American Medical Informatics Association (AMIA) members who belonged to four working groups (WGs): CIS, Evaluation, ELSI, and POI. Other AMIA members joined in significant numbers. It immediately became a major forum for discussing medical informatics, informatics policies, and discussion of the purpose of AMIA itself. The listserv membership approximates 25% of AMIA’s members and has generated over 6,000 posts. We report on a survey of the listserv’s members: what members think about the listserv; what participants want for medical informatics; how they think those goals should be achieved, and what AMIA’s role should be in this process. The listserv provides vital signs about AMIA and hopes for informatics. We combine qualitative analysis of members’ comments and responses about the listserv using ATLAS.ti qualitative text analysis tool and a word cloud generator.

Introduction

The lifeblood of a membership organization like the American Medical Informatics Association (AMIA) is, not surprisingly, its members. Among the reasons members join is because they value interacting with other members through conferences, journals, committees, working groups, and listservs. Each AMIA Working Group (WG) has its own listserv, through which members exchange information, carry on discussion, post announcements, organize conference sessions, and keep in touch with each other. In addition, one cross-group listserv serves these functions for the large section of the membership interested in implementation issues.

Twelve hundred members from AMIA (~25% of the organization’s members) subscribe to the AMIA “Implementation and Optimization Forum” listserv (hereafter, the “listserv”). This listserv was created in 2012 when multiple cross-postings to the four separate AMIA WGs’ listservs frustrated participants with overlapping and redundant conversations. The postings’ high volume and engaging content motivated the four groups to form a new combined listserv to which all members of the four WGs – the Clinical Information Systems WG, the Evaluation WG, the Ethical, Legal and Social Issues WG, and the People and Organizational Issues WG – were automatically
subscribed. Other AMIA members also were invited to join. By October of 2014 the participants had posted 5392 messages in 824 separate discussion threads. As expected, there is a Pareto-like distribution of members’ postings, i.e., some members post far more frequently than many others.

The lively discussions and frequent interchanges reflect the pulse of AMIA membership. Informaticians from different disciplines, countries, and traditions discuss a range of topics, including implementation experiences, ethical conundrums, change management, organizational issues and politics, regulatory environment, workflow, open source, influential publications, design, and cutting-edge content that has not yet reached medical informatics textbooks or journals. Moreover, the listserv was intended to create synergy and energy across working groups. This has apparently been accomplished. Many from outside of the previous four groups have joined the new listserv.

While listservs are not new, they still create community, enhance communication, enable sharing information and resources, and provide a venue for engaging discussion. [1-9] However, there is little research on how they can be used to support knowledge dissemination and how individual scientists use a listserv for various purposes, including education, conversation, and knowledge dissemination. In December 2014, we surveyed members of the listserv to gain insight into their use of and perceptions about the listserv itself. The survey solicited members’ views on what they most appreciated about this listserv, their suggestions to improve it, and topics they would like to have discussed. We report responses from over 200 AMIA listserv subscribers who responded to fixed-choice and open-ended questions.

We previously reported on some of the listserv content and discussions [10-11] and also have reported some preliminary survey findings from early respondents that focused on the fixed-choice survey responses [12]. This paper complements that work in that we focus on the free text comments. They indicate the reasons AMIA members find the listserv of considerable value and ways they think it might be improved. Our survey results indicate how a significant proportion of AMIA members interact, and would like to interact, suggesting ways AMIA can augment on-line member benefits and services. They also point to topics members consider important for improving implementation of health information technologies (HIT), and, ultimately, health care, that AMIA members, and AMIA as an organization, can address. The survey results indicate what members consider important, showing the vitality of both the listserv and of members’ interest in expanding both its functionality and discussion. The survey, therefore, as well as the listserv, measures some of AMIA’s “vital signs.”

Methods

Three methods were used. First, a survey was conceived to gain insight into AMIA members’ perceptions about the listserv. The authors developed the survey instrument between July and September 2014. The survey was pre-tested starting in mid-September 2014 with a group of 50 volunteer AMIA members who were asked to answer the questions and provide suggestions for revising them. Based on the feedback from the pretest, the authors then formulated and iteratively refined the questions via pretesting with 17 of the volunteers. The final version of the survey had 8 questions that ranged from how often people read and make listserv postings to what they like and dislike about the listserv, as well as what they would like to see improved. With the exception of Question 8, all questions were fixed-choice, and Questions 4, 5, 6, and 8 also asked people for free text comments. The fixed choice questions were not limiting in choices, so a respondent could select as many choices as they wish. The survey was IRB approved and respondents were offered anonymity. The survey was officially launched on December 2, 2014 via SurveyMonkey®. [13] All 1200 Forum members were invited to participate and, after several reminders, by February 1, 2015, 223 people responded to the survey.

Second, the numerical data was analyzed by SurveyMonkey® using descriptive statistics. Third, the free text comments were exported into ATLAS.ti [14] and analyzed to identify themes related to the numerical data. Three of the authors (BK, KR, and CK) used ATLAS.ti to independently code the free text data for Questions 4, 5, and 8, focusing on what people most appreciated about the listserv and their suggestions for improvement. After completing their independent coding the three authors then integrated their analyses and used an iterative process to develop thematic consensus. Free text quotations are used throughout the findings to emphasize or further illustrate certain themes. The other authors (MA, SA, and RK) also analyzed the free text data from Questions 4, 5 and 8 questions to identify respondents’ suggestions for AMIA. The four comments to Question 6, which asked participants if they referred others to the listserv, were not analyzed and are not germane to this paper.
For further clarity, a word cloud provided a visual display of each of the Questions 4, 5, and 8 free text comments. For this analysis, the word cloud only gives size as a visualization of word frequency, not necessarily the importance of the words, a task that falls to the authors for the interpretation. [15]

Results

The response rate for the survey was 18.5% and the final number of participants was 223. The findings are structured according to the data from questions 4, 5, and 8. For each question we briefly summarize the fixed choice responses, rounded to the nearest integer, then describe the themes from the free text responses, and conclude with a Word Cloud that illustrates the free text responses. When we describe the themes from the free text data we use quotations from the data to emphasize or clarify certain themes.

The first question about the listserv asked “What aspects of the listserv do you most appreciate?” Two hundred and twenty three fixed choice responses were received and were overwhelmingly positive: Approximately 60% of the respondents said the value of the listserv is in “interesting comments from colleagues (64%), “different perspectives” (64%), and “learning about important issues” (59%). Others commented on “the relative absence of political discourse” and the benefit of access to an international perspective. Moreover, in answer to Question 6, 52% said they have recommended the listserv to others.

While many respondents favored leaving the listserv as is (27% said “don't change a thing”), the survey responses also indicated that some see room to improve. The answers to a subsequent open-ended question regarding how the listserv could be improved revealed three broad categories of recommendations: content of discussions, social rules, and technological improvements. We later discuss each of these in turn.

The free text (additional) comments expanded on the themes noted above. Twenty-one free text comments were received for the question of what respondents most appreciated. Seven of the 21 comments from this question highlighted how much people appreciated getting different perspectives from reading listserv postings. They liked reading different points of view and “the chance to discuss possible different way of doing things….” Perhaps most important is that the interdisciplinary perspective of the listserv was something that both clinical and non-clinical participants appreciated.

(Example response) As a clinician, it shows me how non-clinicians and academic personalities view issues...and helps me understand what physician informatics leaders are thinking.

Participants also reported reading real life experiences from actual HIT implementations. Some noted “real world experiences…show how the grass is not always greener across the pond.”

The second most appreciated value of the listserv was keeping readers informed of current developments. Four comments: new publications, “what’s hot,” upcoming webinars, AMIA activities, and instant dispersion of information.
Figure 1: Word cloud reflecting responses (both fixed choice and free text) for Question 4: “What aspects of the listserv do you most appreciate?”

Fig. 1’s word cloud reflects the sentiments expressed in both the fixed choices and free text comments about the aspects most appreciated in the listserv. Predominant terms are generally positive and welcoming: “learning,” “new developments,” “funny,” “humor,” “informs me,” “new developments,” “experiences” (explained, related), “value,” “discussions of problems,” “helps me understand,” “shows me,” (discusses) “reasons for actions,” “studies,” “publications,” “rapid information,” and “collaboration.”

In tabular form, the most common responses were:

<table>
<thead>
<tr>
<th>What aspects of the listserv do you most appreciate?</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Interesting comments from colleagues</td>
<td>64%</td>
</tr>
<tr>
<td>Differing perspectives</td>
<td>63</td>
</tr>
<tr>
<td>Learning about important issues</td>
<td>58</td>
</tr>
<tr>
<td>Wide range of experiences reflected in the posts</td>
<td>56</td>
</tr>
<tr>
<td>Helps me stay abreast of developments</td>
<td>53</td>
</tr>
<tr>
<td>I appreciated the controversies reflected in the posts</td>
<td>52</td>
</tr>
<tr>
<td>Help me stay connected to the discipline</td>
<td>47</td>
</tr>
<tr>
<td>Informs me about new studies and publications</td>
<td>41</td>
</tr>
<tr>
<td>Some are well written</td>
<td>36</td>
</tr>
<tr>
<td>Some are funny</td>
<td>34</td>
</tr>
<tr>
<td>Help me learn about other participants</td>
<td>31</td>
</tr>
<tr>
<td>Some are written by people I follow/read</td>
<td>27</td>
</tr>
<tr>
<td>Help me stay connected to colleagues</td>
<td>22</td>
</tr>
<tr>
<td>All of the above</td>
<td>26</td>
</tr>
</tbody>
</table>

Note: Respondents could choose multiple items.
The next question asked for “Suggestions for improving the listserv” (N=209). The most popular fixed choice response requested that we “Ask people (politely) to make the headers reflect the actual topic,” which was chosen by 57% of respondents. As noted above, however, 27% of respondents selected “don’t change a thing” as a suggestion for improvement.

Of the 19 free text responses, some addressed technical issues, such as links not loading from postings, but most primarily referred to social aspects of the listserv: certain people dominating the discussions, some “rants” and “rock throwing,” and too many interpersonal discussions.

People’s suggestions to improve the social issues were varied. Many loved the freedom and spontaneity of the listserv, but some felt this led to tangents of limited interest. Respondents were equally divided on whether the list should be moderated. For some, the volume, repetition, and negativity made the listserv less attractive, while others thought it is a “must read.” While some wished for more intervention and listserv moderating (12 of the comments mentioned this) to improve netiquette and reduce the “bashing” and “beating the same drum over and over again,” they were equally split (11 of the comments) because others liked the “spontaneity.” Many welcomed new perspectives and the “frank exchange” of problems being exposed via the “savvy” postings.

(Example response) A few suggested the use of “providing polite training for those with knowledge gaps of how to respond to single individual vs all.”

Figure 2: Word cloud reflecting comments (free text) in response to question 5, about improving the listserv.

The word cloud for this question (Fig. 2) reflects both the joy and frustration of the readers. On the joy side – the predominant side – we see terms like: “useful,” “original,” “great sprit(ed),” “constructive,” “discussions,” “conversations,” and “communications.” On the “frustrated” side, we find terms such as: “header” (as in fails to keep headers up to date), “attacks,” “personal,” “soapbox,” and “personal” (as in responses to personal comments or the rounds of congratulatory messages that some experience as clogging the list).

Finally we inquired: “What else would you like to tell us? For example, some have suggested we encourage more discussions of: cyber security of EHRs and PHRs; the role of the CMIO; specific concerns about the cost of
healthcare IT (both initial software and the cost of implementation); inclusion of clinicians at key decision points in HIT design and implementation.” This question had only free text responses and 97 were received. Suggested topics included: improving EHRs and their implementation, usability or lack of it, HIT errors, standards (e.g., interoperability standards), the continuum of change (from talk to change), liability for lost data, HIT risks, change management, measurement and improvement of quality and safety, the barriers between IT and clinical and fiscal or monetary influences, gaps between technical people and healthcare professionals, and the topics listed in the question itself. Some suggested a topic of the month, while others liked that topics evolved, responded to the news, and reflected member interests and issues. A few commented that they liked the political discussion and ways to effect change, while others wanted less or no political discussion. Some also believe more presence of vendors as well as healthcare students who are currently trained using EMR systems in the listserv could be helpful. Overall, the freedom and spontaneity of the listserv was appreciated by some users but disliked by others who believe it leads to less meaningful dialogue.

(Example response) The beauty of the listserv is the spontaneous presentation of topics represented based on what's relevant to all of us in the field and front lines. Attempting to influence the content detracts from this. You always have the choice to ignore a post or present a topic.

However, others would prefer to have set topics to focus the discussion.

(Example response) I wish it could somehow be focused, as the volume is far (!) too great. Really limit it to implementation and optimization issues, not every topic under the Sun. Some people seem to spend virtually all their time posting to the list...a more circumspect approach would be better.

More specifically, the diversity of participants meant that there were different ideas for selected topics. One idea was topics with a greater academic focus such as social or economic theories or foundations in applied informatics research. Another suggested topic was more perspectives for CMIOs. Other requests were for either more moderator involvement (help focus topics, encourage polite behavior) or for less moderator involvement.

Several commented on the volume of postings and the chaotic or vestigial headers (they don’t change when topics change). Some asked for summaries, search and visualization tools, and more digest options. Some wanted previous messages to be repeated, finding it easier to make sense of the thread, while others “hated” it and wanted them either eliminated or “snipped.” Others argued for mobile apps and greater access via the AMIA website.
In Fig. 3 we see a reflection of the themes noted above: on the one hand, an appreciation of the freedom and evolving nature of the listserv, on the other, a desire to focus and control the discussion.

**Discussion**

The listserv evolved from the intersection of several AMIA WGs. The informality continues in this combined service and is clearly appreciated by members who enjoy the natural development and evolution of topic threads. Nonetheless, some members indicated that they missed specific (types of) HIT implementation topics or that some topics were not discussed adequately and suggested that the list moderator could ensure that these topics were raised, for example by posting a “topic of the month.” Recurring themes in the suggested topics were gaps in applied informatics research, theories and frameworks, standards, interoperability, EHRs/PHRs, and the role of the CMIO. Other members suggested not trying to influence content because one of the listserv’s benefits is its spontaneity of posting topics and the evolution that threads take.

Several respondents felt that just exchanging ideas on the listserv is not enough. In the free text, respondents indicated that the list members should try to move from merely discussing barriers or problems to finding solutions (“moving from talk to change”). Suggestions for doing so included AMIA members collaborating to develop a medical record system, picking up a Quality Improvement initiative or study, and trying to resolve the seemingly ever-present gap between professionals (e.g., technicians versus physicians).

**Social rules**

Some respondents further indicated that the added value of the listserv can become clouded by the amount or nature of posts from a few users with personal agendas. They suggested that a moderator or others should say what types of posts are appropriate for a listserv and what types of posts might be more appropriate via, for example, personal email, blog, or tweet. Moreover, the respondents suggested a level of mutual respect – recognizing that others may have a different opinion from one's own and that some members might be more critical than others in their...
comments. That said, there were far fewer negative comments than positive, and respondents split on the desirability of a moderator.

Additional social rules included awareness of the tone of messages, indicated in statements such as “I really don’t appreciate the sarcastic humor,” and the need to discourage “rants” or “beating the same drum.” Some respondents were careful to say that they do not explicitly favor censorship except in extreme cases where the postings are detrimental to what readers can get out of the list or involve personal attacks.

To address these concerns, some members suggested solutions such as establishing rules of engagement for newcomers to the list or distributing a regular operational message to the list with a policy statement. An additional suggestion was to offer training (to new and existing members) regarding proper listserv etiquette. There were also suggestions to let the moderator play a stronger role, both in the discussions and behind the scenes by explicitly discussing problematic behavior with the person posting to the listserv.

**Technology**

Some responses also indicated that the flurry of discussion that naturally evolves should be balanced against email overload. These respondents therefore asked AMIA administrators to examine technical solutions to help reduce the volume of messages in people’s inboxes. Some members favor a periodic summary of trending topics.

One of the most prominent comments involved threading and individual posting behavior vis-a-vis thread/header hygiene. Few pay attention to the headers even when the topic shifts. Fifty-seven percent indicated that as discussion evolves headers should be adjusted to reflect the actual topic. However, respondents also indicated that this is sometimes difficult within the current listserv software and ease of use with headers. They ask for personal accountability and/or improvement to the current email interface.

**Conclusions**

The implementation listserv, within less than a year, became a major forum for discussing medical informatics, informatics policies, and of the purpose of AMIA itself. At the time of the survey, the listserv was subscribed by 25% of AMIA’s members and, by March 12, 2015 had generated over 6,000 posts. In summary, the survey results analysis serves to advance research evidence about knowledge dissemination, particularly how members interact and what they consider priorities for the topics under discussion and for the listserv itself. This report has relevance both to AMIA and to a more general audience interested in sociotechnical aspects of online knowledge transfer, engagement among community-of-practice members, organizational use of IT, and implementation and optimization issues in deploying and using IT.

For the respondent sample size, 18.2%, we make no claims of representativeness – in part because we turned off the email tracing function as part of our promise of confidentiality and anonymity, and thus cannot report on the identities of individual respondents. We also did not collect demographic data out of the same concern, as we personally know respondents we might have been able to identify. We do know that AMIA membership includes clinicians, researchers, HIT practitioners, and vendors, no doubt reflected in the pool of subscribers to this listserv. In addition, we may assume that those more engaged in the listserv’s discussions were more likely to respond to the survey. We note, however, that some respondents said they seldom if ever contributed to the listserv, which may indicate at least some inclusion of "lurkers."

Of course, the usual limitations of question wording apply. We also know that some for whom English is not a mother tongue may be especially reluctant to contribute to the listserv. On the other hand, we doubt that many members of the listserv are illiterate, making this concern less relevant.

Next steps will be more in depth studies of the listserv and its data such as social network analysis and studies of specific topics (e.g., data standards, workflow, HIT evaluation, international contexts). Our plan is to engage the listserv participants to form groups corresponding to topics of interest.
References

Public Perspectives of Mobile Phones’ Effects on Healthcare Quality and Medical Data Security and Privacy: A 2-Year Nationwide Survey

Joshua E. Richardson, PhD, MLIS, MS, Jessica S. Ancker, MPH, PhD
Department of Healthcare Policy and Research
Weill Cornell Medical College, New York, NY

Abstract

Given growing interest in mobile phones for health management (mHealth), we surveyed consumer perceptions of mHealth in security, privacy, and healthcare quality using national random-digit-dial telephone surveys in 2013 and 2014. In 2013, 48% thought that using a mobile phone to communicate data with a physician’s electronic health record (EHR) would improve the quality of health care. By 2014, the proportion rose to 57% (p < .001). There were no similar changes in privacy concerns yet nearly two-thirds expressed privacy concerns. In 2013 alone, respondents were more likely to express privacy concerns about medical data on mobile phones than they were to endorse similar concerns with EHRs or health information exchange (HIE). Consumers increasingly believe that mHealth improves healthcare quality, but security and privacy concerns need to be addressed for quality improvement to be fully realized.

Introduction

Healthcare researchers and industry leaders perceive mobile phones (mHealth) as having the potential to revolutionize the ways in which consumers self-manage their health and interact with their healthcare providers. It represents an informatics intervention that is fundamentally different from electronic health records (EHRs) or personal health records (PHRs) because of the ability for consumers to capture, process, and share personal health information without being tethered to an Internet browser.

In particular, smartphones and their applications (“apps”) offer novel opportunities for remote patient data gathering such as tracking exercise and sleep patterns or medication adherence. This could be a valuable tool that helps manage chronic disease. It is also encouraging that mHealth is being widely adopted by various consumer groups regardless of race, gender, or socioeconomic status. In fact, there is evidence to suggest that those with lower socioeconomic status may be more willing to accept mHealth than other groups. That may provide a host of new opportunities to engage patients in health care and research. Taking together these trends, mHealth offers exciting opportunities for delivering services in ways that can potentially improve overall healthcare quality. However, along with its rise in availability and use, mHealth may be subject to unintended consequences such as increased threats to data privacy and security.

Informatics researchers and developers are not only applying mHealth strategies to track patients with highly prevalent conditions such as diabetes, but also conditions that may be considered socially sensitive. Examples of those applications include monitoring activities in the context of care for sexually transmitted infections, drug treatment support, mental health care, and delivering healthcare services for patients living with HIV/AIDS. With the continued increase in smartphone adoption, the emergence of wearable devices, and personalized/genetic medicine it is likely that in the near future mHealth data will be collected for a myriad of other sensitive patient conditions in far greater quantities than today. This raises numerous challenges that informaticians will have to address.

One particular challenge is gauging consumer perceptions of mHealth privacy and security with respect to any potential for improvements in healthcare quality. If consumers believe that the gains from mHealth in terms of improved healthcare quality outweigh the costs associated with risks to patient data privacy and security, it is reasonable to assume consumer adoption of mHealth would be broad and therefore the impact would be widely felt. Capturing consumer attitudes will help identify potential opportunities and barriers to mHealth adoption, with implications for mHealth, public policies in the form of new regulations, industry investment and market development, and mHealth research.
Our objective for this study was to gauge consumer perceptions of the effects of mobile devices on medical data security and privacy and healthcare quality.

**Methods**

The Cornell National Social Survey is an annual random-digit-dial telephone survey conducted by Cornell Survey Research Institute. Every year, the sample size of 1000 provides a margin of error of plus or minus 3.1 percentage points. The Cornell University Institutional Review Board approved the study, and respondents provided oral consent.

**Sampling strategy**

The Cornell National Social Survey is conducted annually with a random sample drawn from a dual frame of landline and cell phone numbers in the continental United States not stratified by geographic region, race, or other variables. The proportion of cell phone numbers is calculated from county-level data on prevalence of cell phone-only households. Listed and unlisted numbers are both included in the sample list; known business, disconnected, and non-household numbers are excluded. When the telephone is answered, the adult with the most recent birthday is interviewed to ensure that each adult has equal chance of selection.21

**Survey development and administration**

Researchers from Cornell University and Weill Cornell Medical College submit potential questions that are competitively reviewed by the Cornell Survey Research Institute. Three questions about “cell phones or smartphones” were included for 2 consecutive years: 2013 and 2014. Initial question wording was based on EHR and health information exchange (HIE) questions, which were included within the same survey in 2013 only and are the focus of a separate publication.22 There were slight wording variations based on between-year pilot testing (boldface below) but testers did not note any difficulties with respondent comprehension in either year. In each year one question asked about perceived impact on quality, and two questions asked about perceived impact on privacy and security with regard to sharing health data and storing health data.

**Table 1: 2013 vs. 2014 Mobile Phone Questions**

<table>
<thead>
<tr>
<th>Concept</th>
<th>2013 wording</th>
<th>2014 wording</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Question stem]</td>
<td>If patients used cell phones or smartphones to share personal health data with doctors' electronic medical records, how do you think that would:</td>
<td>If patients used cell phones or smartphones to share personal health data such as cholesterol scores or physical activity levels with doctors' electronic medical records, how do you think that would:</td>
</tr>
</tbody>
</table>
| Effect of data sharing on quality | Affect the quality of medical care?  
  • (5-point Likert scale from “greatly improve it” to “greatly worsen it,” midpoint at “no effect”) | Affect the quality of medical care?  
  • (5-point Likert scale from “greatly improve it” to “greatly worsen it,” midpoint at “no effect”) |
| Effect of data sharing on privacy/security | Affect the privacy and security of medical information?  
  • (5-point Likert scale from “greatly improve it” to “greatly worsen it,” midpoint at “no effect”) | Affect the privacy and security of medical information?  
  • (5-point Likert scale from “greatly improve it” to “greatly worsen it,” midpoint at “no effect”) |
| Effect of data storing on privacy/security | If medical information could be stored electronically on a patient's cell phone or smartphone, how do you think that would affect the privacy and security of medical information? Do you think it would:  
  • (5-point Likert scale from “greatly improve it” to “greatly worsen it,” midpoint at “no effect”) | If medical information in patient records could be stored electronically on a patient's cell phone or smartphone, how do you think that would affect the privacy and security of medical information in patient records? Do you think it would:  
  • (5-point Likert scale from “greatly improve it” to “greatly worsen it,” midpoint at “no effect”) |
Data analysis

Data are summarized with descriptive statistics. Chi-squared tests were used to compare frequencies from year to year. In bivariate analyses, there were small but statistically significant associations between some of the demographic characteristics and the 3 mobile phone questions. For multivariable logistic models (with dichotomized response variables) characteristics significant at .05 were included to determine whether the year-to-year changes remained statistically significant while controlling for participant characteristics. Analyses were performed in SAS (version 9.3).

Results

A total of 1000 respondents were included in 2013 and 2014 with a 3.1% margin of error in each year. The cooperation rate (using the American Association for Public Opinion Research definition of number of eligible participants reached by telephone who participated) was 70% in 2013 and 79.1% in 2014. The samples were somewhat more likely to be white, non-Hispanic, and highly educated than the US population, but were otherwise fairly representative.

Table 1: Participant Demographics in 2013-14

<table>
<thead>
<tr>
<th></th>
<th>2013 N</th>
<th>2013 %</th>
<th>2014 N</th>
<th>2014 %</th>
<th>Total n</th>
<th>Total %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;40</td>
<td>380</td>
<td>38.0</td>
<td>338</td>
<td>33.8</td>
<td>718</td>
<td>35.9</td>
</tr>
<tr>
<td>40-64</td>
<td>431</td>
<td>43.1</td>
<td>443</td>
<td>44.3</td>
<td>874</td>
<td>43.7</td>
</tr>
<tr>
<td>65+</td>
<td>189</td>
<td>18.9</td>
<td>215</td>
<td>21.5</td>
<td>404</td>
<td>20.2</td>
</tr>
<tr>
<td>Female gender</td>
<td>502</td>
<td>50.2</td>
<td>498</td>
<td>49.8</td>
<td>1000</td>
<td>50.0</td>
</tr>
<tr>
<td>Married</td>
<td>564</td>
<td>56.4</td>
<td>536</td>
<td>53.6</td>
<td>1100</td>
<td>55.0</td>
</tr>
<tr>
<td>Children in household</td>
<td>374</td>
<td>37.4</td>
<td>327</td>
<td>32.7</td>
<td>701</td>
<td>35.0</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ High school</td>
<td>248</td>
<td>24.8</td>
<td>229</td>
<td>22.9</td>
<td>477</td>
<td>23.9</td>
</tr>
<tr>
<td>Some college</td>
<td>318</td>
<td>31.8</td>
<td>308</td>
<td>30.8</td>
<td>626</td>
<td>31.3</td>
</tr>
<tr>
<td>College graduate</td>
<td>434</td>
<td>43.4</td>
<td>460</td>
<td>46.0</td>
<td>894</td>
<td>44.7</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>844</td>
<td>84.4</td>
<td>855</td>
<td>85.5</td>
<td>1699</td>
<td>85.0</td>
</tr>
<tr>
<td>Black</td>
<td>122</td>
<td>12.2</td>
<td>98</td>
<td>9.8</td>
<td>220</td>
<td>11.0</td>
</tr>
<tr>
<td>Other</td>
<td>41</td>
<td>4.1</td>
<td>45</td>
<td>4.5</td>
<td>86</td>
<td>4.3</td>
</tr>
<tr>
<td>Born in US</td>
<td>916</td>
<td>91.6</td>
<td>900</td>
<td>90</td>
<td>1816</td>
<td>90.8</td>
</tr>
<tr>
<td>Household income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$50,000</td>
<td>379</td>
<td>37.9</td>
<td>335</td>
<td>33.5</td>
<td>714</td>
<td>35.7</td>
</tr>
<tr>
<td>$50 to &lt; $100,000</td>
<td>343</td>
<td>34.3</td>
<td>363</td>
<td>36.3</td>
<td>706</td>
<td>35.3</td>
</tr>
<tr>
<td>$100,000+</td>
<td>268</td>
<td>26.8</td>
<td>280</td>
<td>28.0</td>
<td>548</td>
<td>27.4</td>
</tr>
<tr>
<td>Employment status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>635</td>
<td>63.5</td>
<td>622</td>
<td>62.2</td>
<td>1257</td>
<td>62.8</td>
</tr>
<tr>
<td>Unemployed</td>
<td>156</td>
<td>15.6</td>
<td>144</td>
<td>14.4</td>
<td>300</td>
<td>15.0</td>
</tr>
<tr>
<td>Retired</td>
<td>149</td>
<td>14.9</td>
<td>184</td>
<td>18.4</td>
<td>333</td>
<td>16.6</td>
</tr>
<tr>
<td>Disabled/Unable to work</td>
<td>60</td>
<td>6.0</td>
<td>50</td>
<td>5.0</td>
<td>110</td>
<td>5.5</td>
</tr>
<tr>
<td>Social views</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liberal</td>
<td>331</td>
<td>33.1</td>
<td>312</td>
<td>31.2</td>
<td>643</td>
<td>32.2</td>
</tr>
<tr>
<td>Moderate</td>
<td>350</td>
<td>35.0</td>
<td>351</td>
<td>35.1</td>
<td>701</td>
<td>35.0</td>
</tr>
<tr>
<td>Conservative</td>
<td>318</td>
<td>31.8</td>
<td>334</td>
<td>33.4</td>
<td>652</td>
<td>32.6</td>
</tr>
</tbody>
</table>

Each year, there was strong support for being able to use mobile phones to share data with a doctor’s EHR. The proportion of respondents who thought this would improve healthcare quality rose from 48% to 57% (p < .001; Figure 1). However, each year the level of privacy and security concerns was also very high. About three-quarters of individuals believed that storing medical information on the phone would threaten privacy and security (74% in 2013, 75% in 2014, p = 0.64). About two-thirds thought that sharing data between a mobile device and a doctor’s EHR would threaten privacy and security (69% in 2013, 67% in 2014, p = .45)
Each year, several characteristics had modest but statistically significant bivariate associations with the 3 mobile phone questions. These are listed in Tables 3-5 and were included in multivariable models. The multivariable models confirmed that there was a statistically significant increase in the proportion that believed mobile phone-EHR communication would improve healthcare quality, with no year-to-year differences in the other 2 mobile phone questions. Younger individuals, those with more education, and nonwhites were significantly more likely to endorse the belief that mobile phone-EHR communication would improve quality. Non-Hispanics were significantly more likely to express privacy concerns about storing data on mobile phones.

Table 2: Associations with Belief that Mobile Phone-EHR Communication would Improve Quality

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>AOR*</th>
<th>95% Wald Confidence Limits</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year: 2014 vs 2013</td>
<td>1.46</td>
<td>1.22 1.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age: 40-64 vs &lt;40</td>
<td>0.73</td>
<td>0.60 0.90</td>
<td>.70</td>
</tr>
<tr>
<td>Age: 65+ vs &lt;40</td>
<td>0.58</td>
<td>0.45 0.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Education: &lt;= High school vs College graduate+</td>
<td>0.67</td>
<td>0.53 0.85</td>
<td>.08</td>
</tr>
<tr>
<td>Education: Some college vs College graduate+</td>
<td>0.67</td>
<td>0.54 0.83</td>
<td>.046</td>
</tr>
<tr>
<td>White race vs Non-white race</td>
<td>0.76</td>
<td>0.59 0.99</td>
<td>.04</td>
</tr>
<tr>
<td>Social views: Liberal vs Moderate</td>
<td>1.18</td>
<td>0.94 1.48</td>
<td>.08</td>
</tr>
<tr>
<td>Social views: Conservative vs Moderate</td>
<td>0.97</td>
<td>0.78 1.21</td>
<td>.26</td>
</tr>
</tbody>
</table>

*AOR, Adjusted Odds Ratio

Table 3: Associations with Belief that Mobile Phone-EHR Communication would Worsen Privacy/Security

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>AOR*</th>
<th>95% Wald Confidence Limits</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year: 2014 vs 2013</td>
<td>1.08</td>
<td>0.89 1.30</td>
<td>.46</td>
</tr>
<tr>
<td>Gender: F vs M</td>
<td>0.83</td>
<td>0.68 1.01</td>
<td>.06</td>
</tr>
<tr>
<td>Hispanic: Yes vs No</td>
<td>1.64</td>
<td>1.20 2.25</td>
<td>.002</td>
</tr>
<tr>
<td>Employment: Disabled vs Employed</td>
<td>1.25</td>
<td>0.82 1.89</td>
<td>.66</td>
</tr>
<tr>
<td>Employment: Retired vs Employed</td>
<td>1.44</td>
<td>1.11 1.87</td>
<td>.046</td>
</tr>
<tr>
<td>Employment: Unemployed vs Employed</td>
<td>1.02</td>
<td>0.77 1.35</td>
<td>.24</td>
</tr>
</tbody>
</table>

*AOR, Adjusted Odds Ratio
<table>
<thead>
<tr>
<th>Variable</th>
<th>AOR*</th>
<th>95% Wald Confidence Limits</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year: 2014 vs 2013</td>
<td>0.96</td>
<td>0.79</td>
<td>1.18</td>
</tr>
<tr>
<td>Born in US: Yes vs No</td>
<td>0.76</td>
<td>0.54</td>
<td>1.06</td>
</tr>
<tr>
<td>Education: &lt;= High school vs College graduate +</td>
<td>1.00</td>
<td>0.78</td>
<td>1.29</td>
</tr>
<tr>
<td>Education: Some college vs College graduate +</td>
<td>0.79</td>
<td>0.62</td>
<td>1.01</td>
</tr>
</tbody>
</table>

*AOR, Adjusted Odds Ratio

In 2013 only, participant opinions about mHealth could be compared to their opinions about physician use of EHRs and HIE. In this year, respondents were markedly more likely to believe that health care quality would be linked to physicians using EHRs or HIE (61% and 74%, respectively) than to patients using mobile devices to communicate with EHRs (48%). They were also markedly more likely to express concern about worsened privacy and security with storing data on mobile phones and mobile phone-EHR communication (74% and 69% respectively) than with EHRs and physician HIE (41% and 47%).

Respondents who believed that EHRs would improve healthcare quality were nearly 3 times as likely to believe that using a mobile device to share data with an EHR would improve healthcare quality (OR 2.94; p < .001).

**Discussion**

Our nationwide survey found increasing year-over-year consumer sentiment that healthcare quality would improve as a result of using mobile phones to share personal health information with doctors’ EHRs. Nevertheless, participants in each year were highly likely to believe privacy and security would worsen as a result from using mobile devices to store and share their personal health information, with no year-to-year change. Of further note is that privacy and security concerns remained high even though the questions posed arguably benign patient scenarios (cholesterol scores and physical activity levels) as opposed to more socially sensitive scenarios, e.g. sexually transmitted infections. In specific, these results strengthen the argument that consumers place high value on privacy and security of health data on mobile phones. These results in general provide important insights that inform informatics research, healthcare policy, and the growing mHealth industry.

Consumers are increasingly optimistic that mHealth can have a positive affect on healthcare quality. This suggests the possibility that consumers are becoming familiar with the idea that mobile devices can be used for user-specific technologies to support overall health or disease-specific tracking: examples include activity monitoring, calorie tracking, and communicating with doctors. The fact that there were no increased concerns about privacy and security may indicate that consumers are not only becoming more familiar with mHealth, but may be more welcoming of mHealth in their personal care. In addition, given that consumers perceived privacy and security as separate from healthcare quality is in itself an interesting finding, for many might argue that privacy and security of mHealth are important features of overall quality. We hope to delve further into this area so to report if and how consumers may perceive the costs and benefits to quality improvement versus threats to privacy and security, and how that may influence future adoption and use of mHealth. In general, researchers may do well to pose future health data privacy and security questions to consumers from within the context of broader health services so that consumers must weigh the costs and benefits for privacy and security.

Despite the generally positive perception of the potential for mHealth, our 2013 comparison found that consumers were more likely to believe that physician-oriented information technologies (EHRs and EHRs with HIE) would improve healthcare quality. It will be interesting to see if consumer perceptions of mHealth and healthcare quality rise to the levels of EHRs and HIE alone in the years ahead.

A large majority of consumers consistently reported that using mHealth to store and share personal data would worsen privacy and security. Researchers, policy makers, and mHealth developers should take note of this result as it indicates possible concern within the general population. Although it did not seem to affect perceptions about how mHealth improving healthcare quality in this study, consumers’ concern could manifest itself in other ways that impact future trends in adoption and use. Also, consumer groups may find cause for concerns in these findings because an increasing amount of informatics research is demonstrating that mobile devices have multiple security threat vectors including data leaks and loss. There may be cause to educate the general public about how to...
securely store and share their protected health information, and for standards bodies to push for more robust mHealth security standards.\textsuperscript{26-28}

Younger individuals and individuals with more education were more likely to endorse the idea that mHealth would improve healthcare quality. These findings may support a view that mHealth is an intervention primarily for the tech savvy. The fact that nonwhites were significantly more likely to endorse that data sharing would improve quality is consistent with research outside of health care showing that nonwhites and those with lower socioeconomic status are just as likely to use smartphones to access the Internet as are whites.\textsuperscript{29} We therefore believe that these findings support future efforts to understand the mHealth user population and how mHealth may cut across traditional lines of race and class. Our finding that non-Hispanics were significantly more likely to express privacy concerns about storing data on mobile phones supports this.

The following are limitations to this study. Sampling coverage and non-response bias are known limitations to random-digit dial surveys. We did not apply survey weights in order to match demographic characteristics of the country because the sampling strategy was random. As a result our sample was fairly representative in gender, age, and employment, but more educated and affluent, more likely to be white, and less likely to be Hispanic than the national population. The survey questions were not separately validated; however, the survey developers piloted the questions for cognitive comprehension and the interviewers did not report any respondent difficulties with the questions. The size and scope of the survey limited our ability to probe particular attitudes regarding privacy and security or mobile devices; we hope to investigate those issues in future work. The study design was cross-sectional, and association does not imply causation.

**Conclusion**

The growing availability of mHealth offers consumers new and exciting opportunities to improve healthcare quality but, as with any informatics intervention, improvements in quality need to be appropriately balanced against potential threats to data privacy and security. Consumers are one of the stakeholder groups, perhaps the key stakeholder group, for determining that balance between potential quality improvements and potential threats to privacy and security. Our 2-year survey found consumers increasingly perceived that using mHealth to share patient data with doctors would improve healthcare quality. Although consumers did not exhibit a year-over-year increase in privacy and security concerns, consumers have strong concerns nonetheless. These findings add encouragement to mHealth advocates but also support the need for informatics researchers, policy-makers, and mHealth developers to improve mHealth privacy and security.

**Acknowledgements**

The Cornell National Social Survey is funded by the Office of the Provost of Cornell University. Jessica Ancker is supported by K01 HS 021531 from the Agency for Healthcare Research and Quality. We would like to acknowledge the Cornell Survey Research Institute for their efforts.

**References**


Automatic Extraction and Post-coordination of Spatial Relations in Consumer Language

Kirk Roberts, PhD, Laritza Rodriguez, MD, PhD, Sonya E. Shooshan, MLS,
Dina Demner-Fushman, MD, PhD
National Library of Medicine, Bethesda, MD

Abstract

To incorporate ontological concepts in natural language processing (NLP) it is often necessary to combine simple concepts into complex concepts (post-coordination). This is especially true in consumer language, where a more limited vocabulary forces consumers to utilize highly productive language that is almost impossible to pre-coordinate in an ontology. Our work focuses on recognizing an important case for post-coordination in natural language: spatial relations between disorders and anatomical structures. Consumers typically utilize such spatial relations when describing symptoms. We describe an annotated corpus of 2,000 sentences with 1,300 spatial relations, and a second corpus of 500 of these relations manually normalized to UMLS concepts. We use machine learning techniques to recognize these relations, obtaining good performance. Further, we experiment with methods to normalize the relations to an existing ontology. This two-step process is analogous to the combination of concept recognition and normalization, and achieves comparable results.

Introduction

Medical ontologies are fundamentally limited in the number of concepts and terms they contain. The range of potential problems, treatments, tests, biological processes, and other concepts encountered in medicine is far too vast to encode directly as concepts in an ontology. Instead, ontologies generally rely on attributes and relations to handle on-the-fly concept creation. Further, for a given concept, it is often impossible to enumerate all the textual strings that can be used to refer to that concept in natural language. The ontological solution for this problem is known as post-coordination, where simple concepts are composed (or coordinated) into a more complex concept. Making full use of an ontology when interpreting unstructured text therefore necessitates natural language processing (NLP) techniques.

The need to perform post-coordination is particularly strong in consumer language. Non-experts lack the breadth of vocabulary that experts often use to describe medical concepts. Instead, consumers often rely on lengthy descriptions with a reduced vocabulary when describing symptoms or conditions. This results in an increased need for post-coordination. Consider the following three examples of consumer language:

1. I am experiencing pain in my left leg.
2. They measured his blood pressure and found it to be above acceptable levels.
3. During the accident I sustained an injury to the back of my head and neck.

The first example corresponds to the UMLS concept Pain in left leg (C0564822 in UMLS 2014AB), yet the presence of the word “my” in the middle of the phrase prevents the concept from being recognized by a simple lexicon lookup. Instead, the concept Pain (C0030193) and Left leg (C0230443) need to be composed to find the pre-coordinated concept in UMLS. In the second example, the patient has Hypertension (C0020538), but this has to be inferred from the result of the high measurement. This is an example of where consumer language would likely differ from that of medical professionals, who might simply state, “He has hypertension.” In the third example, no concept in UMLS corresponds to the problem “injury to the back of the head and neck”. Instead, the problem concept Injury (C3263722), the direction Back (C0205095), and anatomical locations Head (C0018670) and Neck (C0027530) must be post-coordinated into an on-the-fly concept.

Unfortunately, short of complete natural language understanding, there is no general-purpose NLP technique for performing post-coordination. Rather, methods must be used that target specific domains or linguistic phenomena. In this paper, we limit our scope to a vital phenomena for understanding consumer-described symptoms and conditions: the spatial relationships between disorders and anatomical locations. These are generally expressed in one of two ways: (i) a noun compound (e.g., arm pain), or (ii) a grammatical relation between an indicator term such as a spatial
preposition (e.g., *in*, *on*, *at*) and two or more concepts. While noun compounds describing spatial relationships are quite common, they are often pre-coordinated in an ontology and are easier to automatically recognize due to their contiguous nature. The grammatical relations describing spatial relationships are more likely to require special handling and are more difficult to automatically recognize, and hence are the focus of this work. Examples (1) and (3) above demonstrate the second type of spatial relation. While Example (2) contains a spatial relation (*above*), it does not describe the relationship between the concepts of interest here and is therefore outside the scope of this paper.

In this work, we describe both manually annotated datasets and automatic NLP methods to extract disorder-anatomy relations and normalize them to their appropriate UMLS CUIs. Specifically, the contributions of this paper are:

1. A manually annotated dataset based on the Spatial Role Labeling (SpRL) schema\[1\] to extract spatial relations between disorders and anatomical locations.
3. A manually annotated dataset normalizing the extracted spatial relations to UMLS CUIs.
4. An evaluation of several existing automatic methods for normalizing text to CUIs, customized to handle the extracted spatial relations.
5. A discussion of the fundamental limitations to this task based on error analysis of the above automatic methods.

Further, both annotated datasets are being made publicly available via the National Library of Medicine website.

**Background**

While not as well studied as temporal language in medical text, spatial language—especially spatial relations—has nonetheless received considerable attention. Rindflesch et al.\[2\] describes how syntax relationships are crucial to the semantic interpretation of anatomical relationships in cardiac catheterization reports. In particular, they focus on hand-crafted rules for recognizing arterial branching relations as well as the locations of stenosis. Their method relies heavily on the structured data source University of Washington Digital Anatomist\[3\] (UWDA), part of the Foundational Model of Anatomy\[4\] (FMA). Closer to our approach are methods that utilize machine learning (ML) to relate disorders with their anatomical locations. Roberts et al.\[5\] find relations between an inflammation term and its anatomical location within radiology reports. Their method is able to recognize a single layer of relation nesting—that is, when two spatial relations combine into a single conceptual relation (e.g., “*inflammation on the wall of the gallbladder*”). Dligach et al.\[6\] similarly recognize anatomical locations for disorders. Their method is designed to operate on any type of disorder, but they do not recognize nested relations. So in the phrase “*skin tumor removed from behind his left ear*”, the anatomical location for the tumor would be the entire phrase “*behind his left ear*”. The lack of nesting can present several difficulties, however. From a practical perspective, longer phrases are more difficult to recognize automatically. From a linguistic perspective, the phrase still requires further semantic interpretation. In contrast to these two ML methods, our approach can operate on any disorder term and places no restriction on the depth of nesting, and then goes a step further by providing a semantic interpretation. Thus the method could understand a complex spatial phrase such as “*skin tumor on the side of the elbow of his left arm*”.

Also, unlike any of the previous approaches, our focus is on consumer language instead of clinical language. Consumer language has been explored in the past as a means of supporting health information seeking\[7,8,9\] and making medical terms more comprehensible\[10\] for consumers. This work fits more in the former category, where a semantic understanding of consumer language enables systems that connect consumers with health information resources. In previous work with consumer language, we have focused on co-reference\[11\] and question classification\[12,13\].

Outside of medicine, spatial relations have received more attention. Several schemas have been proposed for natural language, including SpatialML\[14\], SpRL\[1\], and ISO-Space\[15\]. Of these, ISO-Space is the most recent and likely the best developed, especially for representing highly-specified geographic descriptions. For under-specified relations, however, ISO-Space and SpRL are largely interchangeable. For our purposes here, therefore, we use SpRL due to its relative simplicity in the knowledge that our annotations could easily be transformed to ISO-Space in the future should the need arise. SpRL was utilized in two SemEval tasks, in 2012\[16\] and 2013\[17\], while a hybrid SpRL/ISO-Space representation was used in the SemEval 2014 SpaceEval task\[18\]. Successful approaches to these tasks have combined supervised ML, syntactic parsing, and integration of real-world knowledge\[19\].
Post-coordination is generally studied in the context of ontologies and medical coding. For instance, Oniki et al.\cite{20} discuss best practices for pre- and post-coordinating concepts when creating a structured knowledge source based on Clinical Element Models (CEMs). Conversely, Dhombres et al.\cite{21} employ post-coordination when utilizing a structured knowledge source, SNOMED CT, to extend its phenotype coverage. The best known NLP system to perform post-coordination is SemRep\cite{22}, which utilizes MetaMap’s\cite{23} option to identify and normalize short phrases. SemRep then identifies relations (e.g., treatment for, location of) that in many cases can be thought of as post-coordinations. However, SemRep does not, nor does any NLP application we are aware of, attempt to normalize these relations back into an ontology to determine if there is an equivalent pre-coordinated concept as described in this paper.

Instead, the most similar NLP task to the one presented in this paper is the combination of concept recognition and concept normalization.\cite{24,25} In this comparison, concept recognition would be analogous to SpRL relation extraction, while concept normalization would be analogous to relation normalization. In theory, an approach similar to concept recognition/normalization could be used here, but would likely perform poorly for two reasons: (a) concept recognizers typically use sequential classifiers, which perform poorly with long concept spans such as those in Examples (1)-(3) above, and (b) concept normalizers attempt to use all the words in the concept, but the relation structure recognizes that some of the words aren’t relevant (e.g., “my” in Example (1)). In the Discussion, we provide some insights as to how comparable our results are to the state-of-the-art in concept recognition/normalization.

Methods

We begin by describing the two sets of annotated data. Next, we describe the process of automatically extracting spatial relations from text. Finally, we describe how these relations are automatically normalized to UMLS concepts.

A. Data

In this section, we describe two different manually annotated data sets: (1) a set of spatial relations in natural language text, and (2) a set of normalizations from the extracted spatial relations to UMLS concepts. Both datasets are publicly available from the U.S. National Library of Medicine (NLM) website.\cite{26}

Spatial Relations: To gather a set of consumer-written texts likely to contain spatial relations, we started with a large set of emails and online form requests sent to the NLM customer service team. Every year, NLM receives over 40,000 such requests, several thousand of which are manually classified by NLM staff as consumer health questions pertaining to diseases, conditions, and therapies. From these, we extracted 1,976 sentences containing (1) at least one term from UMLS in the DISORDER semantic group,\cite{26} (2) at least one term from UMLS in the ANATOMY semantic group, and (3) a preposition between one UMLS term of each type. Next, two medical experts—an MD/PhD (LR) and medical librarian (SS)—double-annotated spatial relations using a simplified Spatial Role Labeling (SpRL) schema.\cite{1} SpRL defines a spatial relation between three main elements:

1. **SpatialIndicator**: The word or phrase (typically a preposition) that acts as a trigger for the spatial relation.
2. **Trajectory**: The object whose spatial position is being described.
3. **Landmark**: The location of the Trajectory.

In our case, the Landmark is almost always an anatomical location from UMLS (hereafter, an Anatomy annotation), while the Trajectory is usually a Disorder but can be an Anatomy as well. Figure 1 shows how two of the example phrases from the Introduction would be annotated in SpRL using Brat.\cite{27} In the simple example in Figure 1(a), the SpatialIndicator “in” connects the Trajectory “pain” with the Landmark “left leg”. In the more complex example in Figure 1(b), there are two connected spatial relations: (i) a Disorder-Anatomy relation connects the Trajectory “injury” to its Landmark “back”, then (ii) an Anatomy-Anatomy relation connects “back”, this time a Trajectory, to two Landmarks, “head” and “neck”. The annotators were allowed to manually add or edit Disorder or Anatomy terms. Approximately 7% of Disorders were manually created, while around 8% of Anatomies were manually created. Since the focus of this paper is relation extraction and normalization, we do not address automatic concept extraction. The UMLS and manual terms are distinguished in the publicly available data, though no distinction is made between them in the experiments below.

---

\[1\] http://lhncbc.nlm.nih.gov/project/consumer-health-question-answering
From the 1,976 sentences, a total of 1,291 SpRL relations were annotated. To assess agreement, we consider one annotator’s labels as the “gold” set and measure the other annotator’s labels against this with $F_1$-measure (if the annotators were switched, $F_1$ would be the same, while precision and recall would be reversed). Using this method, the inter-annotator agreement has an $F_1$ of 88.9 for SPATIALINDICATORS, 81.5 for TRAJECORY, 86.2 for LANDMARKS, and 72.67 for the complete relation (i.e., exact match for every argument). The annotators reported their agreement improving over time, where many of the disagreements were the result of ungrammatical and medically incorrect consumer language.

**Concept Normalization:** Instead of uninterrupted text spans, here we consider relations to refer to concepts. Any text not part of a SPATIALINDICATOR, TRAJECORY, or LANDMARK—such as the word “my” in Example (1)—is not considered part of the concept. Further, we consider the fully-nested relation of all DISORDERS and ANATOMYs connected through some SpRL relation—such as merging the DISORDER-ANATOMY and ANATOMY-ANATOMY relations in Example (3) into one nested spatial relation. The nested structure of these two examples would be:

(1) pain

   in: left leg

(3) injury

   to: back

   of: head

   of: neck

This nested structure is then used to normalize relations to an ontology. We used 514 such nested relations from the SpRL relations above and manually normalized them to UMLS concepts as follows. The same two annotators as above double-annotated both the nested relations as well as parts of the relations. Normalizing parts of a relation to UMLS enables two useful applications: (1) post-coordination of concepts when the full relation has no UMLS concept, and (2) assignment of codes to parts of a larger concept to specify its deep structure. For the two examples above, the annotated items would be:

(1) phrase: pain in left leg

   term: pain

   term: left leg

(3) phrase: injury to back of head and neck

   term: back

   phrase: back of head and neck

   term: head

   term: injury

   term: neck

Here, a “phrase” is the result of a relation, while “term” is an individual DISORDER or ANATOMY. We only distinguish for the purpose of evaluation. For each phrase/term, the UMLS Terminology Services (UTS) API\(^2\) is queried for potential concepts belonging to either the Disorder or Anatomy semantic groups. Both the words and approximate matching functions were used, since in our exploratory experiments both these had better recall than the other matching functions. The top 10 results per matching function were retrieved, de-duplicated, and sorted by CUI instead of relevance rank to reduce biasing the annotators toward choosing the top result. The annotators then chose the correct normalization or indicated that none of the concepts were proper normalizations. Table 1 shows an example of how one phrase and one term were annotated.

There were a total of 1,747 annotated phrases and terms. Around 66% of the annotations had a concept normalization (i.e., a pre-coordinated concept in UMLS). Phrases (relations) were far more likely to not have a valid normalization. 74% of the phrases had no normalization, and thus require post-coordination of their component terms to be properly interpreted. By comparison, only 15% of the individual DISORDER and ANATOMY terms have no normalization. About half of the terms without a normalization are the result of manually creating DISORDER and ANATOMY terms that were not in UMLS. The rest are the result of none of the candidate normalizations being judged valid by the annotators. Inter-annotator agreement was fairly good at 84.8%.

\(^2\)https://uts.nlm.nih.gov/home.html#apidocumentation
B. Spatial Role Extraction

Figure 2 presents the architecture of our ML-based SpRL relation extractor. We utilize a pipeline method where Disorders and Anatomys (from their respective UMLS semantic groups) are pre-annotated. Next, SpatialIndicators are extracted. Then, for each SpatialIndicator, Trajectors and Landmarks are identified. These steps are described in detail below except for the pre-annotation, which is described in the Data section.

For candidate SpatialIndicators, we consider all prepositions, adverbs, and participles. Some light verbs (such as have and get) can be indicators, but not sufficiently often to merit their inclusion here. We use Stanford CoreNLP\cite{28} to perform part-of-speech tagging. The primary goal in candidate selection is recall: incorrect indicators can still be filtered out by the classifier, but missing indicators cannot be recovered. This recall-based technique achieves 98% recall with a precision of 14%. This precision is sufficiently high to ensure the training data is not overly imbalanced, thus giving the machine learning classifier below the opportunity to filter out most of the incorrect candidates. A binary support vector machine (SVM) then classifies candidates as positive or negative. The primary consideration in deciding whether a word is actually a SpatialIndicator is its context, especially its relationship with nearby Disorder and Anatomy terms. With that in mind, the features used by this SVM are shown in Table 2(a).

Since SpatialIndicators are the central element of an SpRL relation, candidate Trajectors and Landmarks are defined in terms of the Disorder/Anatomy concept and the SpatialIndicator (e.g., in Example (3), “head” is a valid Landmark for “of” but not for “to”). For a given SpatialIndicator, all Disorder and Anatomy terms within its sentence are considered as candidate Trajectors and Landmarks (a Disorder is rarely a Landmark, but this is possible). Then, two binary SVMs (one Trajector, one Landmark) classify candidates as positive or negative. The primary consideration for a positive Trajector/Landmark is its relationship with its SpatialIndicator, especially its syntactic relationship (using syntactic dependencies obtained from CoreNLP). With this in mind, the Trajector features are shown in Table 2(b), while the Landmark features are shown in Table 2(c).

<table>
<thead>
<tr>
<th>Phrase or Term</th>
<th>Candidate Concepts</th>
<th>Gold Concept</th>
</tr>
</thead>
</table>

Table 1: Examples of the candidate concepts provided to the annotators for normalization (2nd column), as well as the proper normalization selected by the annotators (3rd column).
Post-processing. We then consider two post-processing modules, one for SpatialIndicators and one for both Trajectors and Landmarks. The NoArgFilter prunes any SpatialIndicators that do not have either a Trajectory or Landmark. Since most SpRL relations have all three elements, this filter is essentially a voting strategy where, if the two argument classifiers agree an indicator has no argument, then it is likely that the indicator classifier erred. Then, the AddCoords heuristic adds a new Trajectory or Landmark if an existing Trajectory or Landmark is separated by a concept of the same type (Disorder/Anatomy) by an and or an or. This helps identify errors where only one item in a coordination (e.g., “head and neck”) is marked as an argument.

C. Spatial Relation Normalization

Given a sentence containing at least one SpRL relation, the same process as described in the Data section is used to obtain the fully-nested relations as well as the sub-parts of the nested relation. For the purpose of this paper, we are not proposing any new methods for normalizing concepts. Instead, we evaluate baseline and existing methods for concept normalization, and customize these methods for the spatial relation structure. Since existing concept normalization methods expect a word sequence instead of a relation, we use all the words in any element of the relation just as for generating concept candidates from UTS. We also filter out any concepts that do not belong to the Disorder or Anatomy semantic groups.

We evaluated the following methods:

1. tfidf pref: (Baseline) Ranks the candidates (returned by the UTS method described in the Data section) using TF-IDF scored cosine similarity between the relation and the concept’s preferred name. Document frequencies are derived from a recent version of PubMed Central (PMC).
2. tfidf all: (Baseline) The same as tfidf pref, but uses every concept name instead of just the preferred name.
3. UTS[exact]: The top result for the UTS exact matching function.
4. UTS[approximate]: Same except for the UTS approximate function.
5. UTS[left trunc]: Same except for the UTS leftTruncation function.
6. UTS[right trunc]: Same except for the UTS rightTruncation function.
7. UTS[words]: Same except for the UTS words function.
8. UTS[norm words]: Same except for the UTS normalizedWords function.
9. UTS[norm str]: Same except for the UTS normalizedString function.
10. UTS[vote]: A combination of the above seven UTS methods using a simple voting strategy. Ties are broken by choosing the concept with the lower CUI number.

Here, the TF-IDF based methods provide simple baselines based on a standard similarity metric. They should work

---
Table 3: SpRL Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Phrases</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>tfidf</td>
<td>61.8</td>
<td>49.8</td>
<td>57.7</td>
</tr>
<tr>
<td>pref</td>
<td>77.0</td>
<td>41.0</td>
<td>59.9</td>
</tr>
<tr>
<td>tfidf</td>
<td>72.6</td>
<td>29.7</td>
<td>43.4</td>
</tr>
<tr>
<td>all</td>
<td>54.4</td>
<td>51.6</td>
<td>57.4</td>
</tr>
<tr>
<td></td>
<td>24.9</td>
<td>20.3</td>
<td>22.6</td>
</tr>
<tr>
<td>tfidf</td>
<td>79.4</td>
<td>79.2</td>
<td>71.6</td>
</tr>
<tr>
<td>pref</td>
<td>79.3</td>
<td>94.3</td>
<td>21.7</td>
</tr>
<tr>
<td>tfidf</td>
<td>79.5</td>
<td>78.7</td>
<td>79.2</td>
</tr>
<tr>
<td>all</td>
<td>63.6</td>
<td>64.0</td>
<td>66.2</td>
</tr>
<tr>
<td>tfidf</td>
<td>80.8</td>
<td>57.6</td>
<td>64.5</td>
</tr>
<tr>
<td>all</td>
<td>28.5</td>
<td>22.0</td>
<td>23.6</td>
</tr>
<tr>
<td>tfidf</td>
<td>75.6</td>
<td>40.9</td>
<td>17.8</td>
</tr>
<tr>
<td>all</td>
<td>28.1</td>
<td>24.6</td>
<td>26.8</td>
</tr>
<tr>
<td>tfidf</td>
<td>78.9</td>
<td>71.7</td>
<td>21.7</td>
</tr>
<tr>
<td>all</td>
<td>70.2</td>
<td>68.2</td>
<td>75.6</td>
</tr>
<tr>
<td>tfidf</td>
<td>84.7</td>
<td>60.7</td>
<td>84.2</td>
</tr>
<tr>
<td>all</td>
<td>49.1</td>
<td>62.5</td>
<td>48.5</td>
</tr>
<tr>
<td>tfidf</td>
<td>73.2</td>
<td>25.0</td>
<td>0.7</td>
</tr>
<tr>
<td>all</td>
<td>73.2</td>
<td>72.0</td>
<td>76.7</td>
</tr>
<tr>
<td>tfidf</td>
<td>89.4</td>
<td>75.3</td>
<td>72.4</td>
</tr>
<tr>
<td>all</td>
<td>73.8</td>
<td>72.4</td>
<td>77.6</td>
</tr>
<tr>
<td>tfidf</td>
<td>81.9</td>
<td>92.5</td>
<td>32.4</td>
</tr>
<tr>
<td>all</td>
<td>73.8</td>
<td>72.4</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Table 4: UMLS Normalization Results

well when relations have a high word overlap with the valid concept, but they do not utilize UMLS in any way. Instead, the UTS API utilizes the structure of UMLS to return candidate normalizations, performing its own internal ranking (though we are not aware of any way in which the internal scores are exposed through the API). In contrast to the systems evaluated in the ShARe/CLEF 2013 task\cite{24}, UTS is not limited simply to DISORDER concepts, and is thus more appropriate for our evaluation, which includes many ANATOMY concepts.

Results

The SpRL method is evaluated using 10-fold cross validation. The results are shown in Table 3. The best SPATIALINDICATOR performance (including the NoArgFilter) has an F₁ of 91.0, while the best TRAJECTOR classifier (without the AddCoord heuristic) has an F₁ of 84.9, and the best LANDMARK classifier (with the AddCoord heuristic) has an F₁ of 86.0. The post-processing NoArgFilter produced a significant gain in precision, from 87.3 to 96.8, for almost no cost to recall. The AddCoord heuristic, however, had only a small effect. In many cases, the classifiers were able to find coordinated TRAJEC'TORS and LANDMARKS without the AddCoord heuristic, which makes sense as the syntactic dependency features should be identical for both sides of the coordination. Overall, these results are quite good, even a bit high compared to similar relation extraction tasks.

The results of the concept normalization methods are shown in Table 4. The baseline methods (tfidf_pref and tfidf_all) performed poorly compared to the best UTS methods. Given that the UTS methods have a better understanding of the UMLS structure and can utilize it for retrieval (e.g., understanding synonymy), they perform quite a bit better and should form the baseline for any future work. However, the individual UTS methods varied significantly, both by method and by the type of concept they were normalizing. UTS[exact] had the best F₁ on individual terms (79.0), but performed quite poorly with the relation phrases (35.3). UTS[approximate] performed decently, but well below the best methods for each concept type. This confirmed our pre-existing observations when experimenting with the methods, as the UTS[approximate] method does a good job of retrieving the correct candidate in the top 10, but a poor job of ranking the correct candidate first. It is therefore good for annotation, but not ideal as a run-time system. The other method used in annotation, UTS[words], performed better, but was still outperformed by other methods. UTS[left_trunc] and UTS[right_trunc] where quite poor in general, followed by UTS[norm_words]. UTS[norm_str] was the most consistent method, and also the best performing method for phrases (73.8). As might be expected when forming an ensemble method from highly noisy constituent methods, UTS[vote] was unable to outperform the best individual methods. For future work, methods should be compared either to UTS[norm_str], or to the use of UTS[exact] for individual terms and UTS[norm_str] for phrases.
Discussion

Using the system described above, many of the detailed concepts found within consumer text can be mapped into an ontology containing either pre-coordinated terms or that allows for spatial post-coordination. Given the informal language often utilized in consumer text, such a method is often necessary to gain a more complete semantic understanding. We now discuss the general types of errors made by the system, how some of these errors illustrate limitations to our approach, and how they motivate future work on this task.

As is typical with consumer-authored text, many of the system’s errors were the result of ungrammatical and error-prone language. This includes misspellings (e.g., “fection” instead of infection, “boh legs” instead of both legs), missing punctuation (e.g., “pain cramping, redness and swelling”), and incorrect tense (e.g., “fingers are stick together by her skin”). These errors are compounded by the fact that many of the consumer information requests were submitted by non-native English speakers, resulting in text that can be difficult to understand (Spatia1Indicators in bold):

(4) i had got sunburn effect before 3 year on my both hands.
(5) i have pain on ma left side under the ribis n difficult on urinating seeming as if threz something blocking is it connected to the kidneys

Other errors were the result of using UMLS directly as a lexicon. While this removes the need to run a noisy concept recognition system, and while the UMLS terms should be far easier to normalize, using UMLS directly introduces noise as well. As stated above, between 7-8% of the Disorder and Anatomy terms needed to be manually created (thus indicating recall issues), but precision issues are an effect as well. For instance, mrs (in the phrase “mrs. p gupta”) and mm (in the phrase “1.3 mm”) are both marked as Disorders, and could easily result in mistaken Trajectors. In many cases, the resulting errors were relatively innocuous, as in the following:

(6) ...she developed water behind her eye balls...

Here, UMLS does not contain the full phrase “eye balls” (though it does contain “eye ball”, but we performed no stemming for relation extraction). Instead, “eye” is selected as the Landmark. Also see Example (8) below.

As with many relation extraction approaches, long distance relations often prove problematic. Consider the example:

(7) The cancer cells are in the fluid, around the lung, and in the trachea lymph nodes...

Here, “cancer cells” is the Trajector for all three Spatia1Indicators, but the classifier fails to identify it as the Trajector for the second and third indicators. The dependency parser should attach all three prepositions to the verb “are”, but instead it attaches “around” and the second “in” to the noun “fluid”. Thus, the dependency paths are atypical for Trajectors, and the classifier misses this argument. Prepositional phrase attachment is a very common source of syntax parsing errors, and the choice of attaching the preposition to the nearer word is typical.

Another common problem is Spatia1Indicators missing either a Trajector or a Landmark. For instance:

(8) Is matastatic non small cell lung cancer able to spread to the spine and brain in as little as 1 month?

Here, the Spatia1Indicator “to” should have a Trajector (“non small cell lung cancer”) and two Landmarks (“spine” and “brain”). While the classifier found both Landmarks, it missed the Trajector. With very few exceptions, all Spatia1Indicators should have a Trajector. But doing this through a post-processing heuristic by simply forcing the nearest concept to be the Trajector hurts performance. In this case, for example, the nearest concept is actually “able”, which is a Finding in UMLS. Instead, in future work, we plan to integrate more intelligent post-processing to enforce constraints such as “Every Spatia1Indicator should have at least 1 Trajector” directly on top of the output of each classifier.

Since the normalization methods evaluated here are either baseline (TF-IDF) or existing (UTS) methods, we omit a detailed error discussion and instead provide some insights on what may enable future methods to improve on normalization performance. The fundamental issue with the evaluated methods is that they consider the concept as a phrase, not as a relation, and thus miss out on potentially useful information. When normalizing a relation, a very similar process to that of post-coordination can be used to identify potentially pre-coordinated concepts. That is, identify the individual argument terms (which almost all of the UTS methods do with higher performance than the
full phrases), then identify valid pre-coordinations. For instance, the phrase “fracture on his thigh bone” results in the Trajector fracture with Landmark thigh bone. The UTSA [words] method identified the concept C0840234 (Fracture of bone in neoplastic disease; pelvic region and thigh), which is overly specific because it indicates the cause of the fracture (a bone tumor). Through relations in UMLS, it can be determined that C0840234 is a pathological fracture and thus too specific, avoiding this incorrect normalization. This type of method would clearly require a significant understanding of the UMLS relation structure and is beyond the scope of this work, but this work does provide the required NLP methods to enable such ontological methods by extracting and structuring spatially related concepts from text. A second, but also fundamental limitation of the normalization approaches used here is the lack of context employed. The context surrounding a relation likely provides valuable clues as to its proper interpretation.

As addressed in the Background section, this normalization process is largely analogous to the types of concept normalization performed in the CLEF task [24], though that task is limited to Disorders. While evaluating an automatic system in that task (e.g., DNorm [23]) would be an unfair comparison based on the type of data, some discussion on the results of the CLEF task and that of this work might provide some useful insights. The CLEF task evaluated systems in two different ways. First, an end-to-end evaluation (“strict”) that required systems to correctly recognize disorder concepts and then normalize the disorders to their UMLS CUIs. Since errors could occur at either the recognition or normalization stage, the results of this evaluation essentially form a lower bound to normalization performance. A second evaluation (“relaxed”) then measured performance of normalization on only those disorders that were correctly recognized. Since the correctly recognized disorders were likely to be shorter terms resembling those in the training data, the relaxed evaluation essentially forms an upper bound to normalization performance. The best system on the strict measure, DNorm, achieved a strict accuracy of 58.9% and a relaxed accuracy of 89.5%. By comparison, in our data the UTSEXACT method has an accuracy of 79.4%. This suggests that the UTS method, though far simpler than DNorm, is not likely to be substantially worse, but would likely nonetheless be outperformed by DNorm if that system were to be tailored to our spatial relations. Due to the complexity of such a project, we leave this idea to future work.

Conclusion

This paper described a method for extracting and normalizing spatial relations between disorders and anatomical structures. While the method might still generalize to physician and other clinical notes, our focus was on consumer language. We have described two new annotated corpora for these tasks: (1) a corpus of spatial relations from consumer health requests, and (2) a set of annotated concept normalizations for many of the relations from the first corpus. A machine learning-based method is used to automatically perform spatial relation extraction on the first corpus. Then, a set of baseline and existing methods are used to automatically perform concept normalization on the second corpus. Both achieve good performance, though we have identified several areas of future work to improve upon these results.

Acknowledgements This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. The authors would also like to thank Parisa Kordjamshidi, Olivier Bodenreider, and Halil Kilicoglu for their valuable input.

References

Automatic Classification of Structured Product Labels for Pregnancy Risk Drug Categories, a Machine Learning Approach

Laritza M. Rodriguez, MD, PhD, Dina Denner Fushman, MD, PhD
Lister Hill National Center for Biomedical Communications, National Library of Medicine,
National Institutes of Health, Bethesda, MD

Abstract

With regular expressions and manual review, 18,342 FDA-approved drug product labels were processed to determine if the five standard pregnancy drug risk categories were mentioned in the label. After excluding 81 drugs with multiple-risk categories, 83% of the labels had a risk category within the text and 17% labels did not. We trained a Sequential Minimal Optimization algorithm on the labels containing pregnancy risk information segmented into standard document sections. For the evaluation of the classifier on the testing set, we used the Micromedex drug risk categories. The precautions section had the best performance for assigning drug risk categories, achieving Accuracy 0.79, Precision 0.66, Recall 0.64 and F1 measure 0.65. Missing pregnancy risk categories could be suggested using machine learning algorithms trained on the existing publicly available pregnancy risk information.

Keywords: machine learning, pregnancy, drug risk, data-mining, knowledge extraction, document classification

Background and Justification

The use of drugs and other substances during pregnancy and lactation can be damaging for the developing embryo and fetus, and is in general discouraged, unless there is a strong medical reason. At any given time over 10 million women are pregnant or lactating in the United States; these women often need to use drugs and other substances while under care of medical providers, likely other than obstetricians. Drug effects on the fetus and embryo vary during different stages of development; the hemodynamic changes in the pregnant women cause changes on the absorption rates, and pharmacologic action of medications and substances. The obstetrical and pharmacological literature information on the effects of drugs on the embryo and the fetus is extensive, however, the information is static, difficult to maintain and not readily available at the point of care. Ideally, this information should be provided in the FDA-approved standard product labels (SPL) made available to the public by the National Library of Medicine (NLM). The SPL for each drug is created by the pharmaceutical manufacturer; it is a requirement of the FDA to provide complete information on use, dose, content, contraindications, side effects and warnings of drugs and chemical products for human and animal consumption. The drug information is stored in a structured standardized format with document sections defined by the LOINC document standard. The LOINC codes and descriptions for the SPL document sections are described in (Table 1).

Table 1: SPL document sections LOINC codes

<table>
<thead>
<tr>
<th>LOINC Code</th>
<th>SPL section name</th>
</tr>
</thead>
<tbody>
<tr>
<td>34066-1</td>
<td>FDA package insert Boxed warning section</td>
</tr>
<tr>
<td>34067-9</td>
<td>FDA package insert Indications and usage section</td>
</tr>
<tr>
<td>34068-7</td>
<td>FDA package insert Dosage and administration section</td>
</tr>
<tr>
<td>34069-5</td>
<td>FDA package insert How supplied section</td>
</tr>
<tr>
<td>34070-3</td>
<td>FDA package insert Contraindications section</td>
</tr>
<tr>
<td>34071-1</td>
<td>FDA package insert Warnings section</td>
</tr>
<tr>
<td>34076-0</td>
<td>FDA package insert Information for patients section</td>
</tr>
<tr>
<td>34084-4</td>
<td>FDA package insert Adverse reactions section</td>
</tr>
<tr>
<td>34088-5</td>
<td>FDA package insert Overdosage section</td>
</tr>
<tr>
<td>34089-3</td>
<td>FDA package insert Description section</td>
</tr>
<tr>
<td>34090-1</td>
<td>FDA package insert Clinical pharmacology section</td>
</tr>
<tr>
<td>38056-8</td>
<td>FDA package insert Structured product labeling supplemental patient material</td>
</tr>
<tr>
<td>42230-3</td>
<td>FDA package insert Structured product labeling patient package insert section</td>
</tr>
<tr>
<td>42231-1</td>
<td>FDA package insert Structured product labeling medguide section</td>
</tr>
</tbody>
</table>
The Code of Federal Regulations (CFR) Title 21 from the Federal Drug Administration (FDA) describes the specific requirements on content and format of labeling for human prescription drug and biological products. It includes specific guidelines for a section on specific populations describing the effects on pregnancy and lactation. The CFR indicates: “the section may be omitted only if the drug is not absorbed systemically and the drug is not known to have a potential for indirect harm to the fetus”. However, the pregnancy section is missing from some of the labels even for those drugs where the information is provided for the same ingredients in other brand names labels. This information is available in proprietary collections, such as Micromedex® Solutions. Micromedex Solutions is an evidenced based clinical resource curated by experienced professionals in the healthcare field. The resource includes the Micromedex Pharmacological Knowledge with a monograph for pharmaceutical products in brand and ingredient forms. The pregnancy risk categories in the Micromedex monographs include curated classes from the FDA classification, the Australian categorization system for prescribing medicines in pregnancy (ACPM), and a simplified classification defined by Micromedex. However, not all documents have categories from all three classification systems: some have only FDA categories and some only the simplified Micromedex categories. For the purposes of this study we considered Micromedex Fetal risk is minimal equivalent to FDA categories A and B, Micromedex Fetal risk cannot be ruled out equivalent to FDA category C, and Micromedex Fetal risk has been demonstrated equivalent to FDA categories D and X. Table 2 summarizes the FDA drug pregnancy drug categories, and the Micromedex classification.

Table 2: Pregnancy risk Categories

<table>
<thead>
<tr>
<th>FDA Pregnancy Risk Drug Categories</th>
<th>Micromedex Pregnancy Category Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Fetal risk is minimal</td>
</tr>
<tr>
<td>Positive evidence of risk: There is positive evidence of human fetal risk based on adverse reaction data from investigational or marketing experience or studies in humans, but potential benefits may warrant use of the drug in pregnant women despite potential risks.</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Fetal risk cannot be ruled out</td>
</tr>
<tr>
<td>Available evidence is inconclusive or is inadequate for determining fetal risk when used in pregnant women or women of childbearing potential. Weigh the potential benefits of drug treatment against potential risks before prescribing this drug during pregnancy.</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Fetal risk has been demonstrated</td>
</tr>
<tr>
<td>Evidence has demonstrated fetal abnormalities or risks when used during pregnancy or in women of childbearing potential. An alternative to this drug should be prescribe during pregnancy or in women of childbearing potential.</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>
The FDA pregnancy risk category is assigned according to scientific evidence of effects of drugs on the developing embryo and fetus based on animal and human studies. Clinical obstetrical pharmacological studies on the effect of drugs focus on limited numbers of drugs, require expert review and are often limited to a small number of drug classes. In recent years there has been an effort by the Advisory Committee of Prescription Medicines (ACPM) in Australia, and the FDA to change the pregnancy risk category classification, and to base the risk classification on human studies and population registries that gather accurate information on the effect of drugs on humans. These efforts will take time and extensive data analysis. For now, the most complete and widely accepted classification remains the FDA risk category. As mentioned above, the information in the SPLs is incomplete and does not include all available drugs. We sought to augment the existing information and assign missing pregnancy risk drug categories to SPLs. To that end, we applied machine learning algorithms to extract knowledge from the free text and leverage the existing categories assigned to some of the documents. Automatic SPL classification has the potential to aid expert groups when studying the effects of drugs belonging to drug classes and subclasses, and could ultimately be applied to aid prescribers at the point of care.

**Materials and Methods**


---

**Figure 1: Example of a SPL header and Warnings Section**

---

1095
We used the following regular expression in Perl to identify and extract the standard pregnancy risk categories from the labels: `/pregnan.{3,50}category[^a-z0-9]+([abcdnx])(?:[^a-z0-9]|$)/i`

We manually checked the pregnancy risk class assigned by regular expressions. Manual checking allowed us to find cases missed by the regular expressions, for example, SPLs in which “pregnancy” is misspelled; the words “pregnancy” and “category” are separated by other text or the category is set off by punctuation.

In the process, we identified two sets of documents: those that had risk categories and those with no mention of pregnancy risk category in the text. We used the standard FDA categories (A, B, C, D, X) and N when no category was mentioned in the text and assigned a risk category class to each document. Some drugs include more than one pregnancy drug category (a drug that is safe in the second trimester of pregnancy might be contraindicated in the third trimester e.g. NSAID), drugs of this type will have mention of two categories in the label. However, the overall total of documents with two categories was only 81; these labels were removed from our data set, since these sparse classes were insufficient to train the classifier. This left us with 15,221 labels in the training set and 3,039 in the testing set with an overall total of 18,260 documents. The class distribution of the labels is summarized in Table 3.

We segmented the documents using the XML section tags and then converted the sections to plain text for the machine learning approach. Each section preserved the original document unique identifier and the pregnancy risk class label. In the testing set, the unknown class label field was replaced with a “?” to comply with WEKA format requirements for unclassified documents.

Machine Learning

Support vector machines are extensively used in biomedical text processing and are known to produce good classification results. Marafino and colleagues demonstrated similar successful results in a different clinical domain using N-grams SVM. Other authors have compared the use of applying different machine learning algorithms Naive Bayes, and Stacking, Boosting and Feature Selection for document classification. We experimented with Naive Bayes and SMO with and without feature selection. SMO had the best cross-validation results, therefore, we used the Sequential Minimal Optimization (SMO) implementation of the method described by Platt, and available in WEKA for our final experiments.

To generate the support vectors we applied the unsupervised 'StringToWordVector' filter to both training and testing sets for all the document sections with the following parameters: lower case and normalization, string delimiters, word counts, N-grams restricted to 3 words per string, no stemmers, no stop word list, preserving 400 most frequent strings per document section. We applied the SMO algorithm classifier to each of the training document sections with 10 fold cross validation to decide which document section had the highest predictive value for the classification task. We used the value of Receiver Operating Curve (ROC) weighted average obtained from the cross validation to select the best performer. The SPL precautions sections had the best performance for document classification for pregnancy risk categories. We applied the resulting classification model trained on the precautions section to the precaution sections of the testing set. We then evaluated the performance on the testing set as described in the next section.

Machine Learning Model Evaluation

To evaluate the performance of the classifier we limited the set to single ingredient drugs using the SPL document mapping to RxNorm drug terminology standard. To do this we used the RxMix tool providing RxCUIs for Standard Clinical Drugs (SCD) as input and limiting the output term type to ingredients (IN). We extracted the ingredients for both the training and the testing sets separately. From the Micromedex Pharmaceutical Knowledgebase monographs we manually extracted the pregnancy risk category for these ingredients.

As previously described, the Micromedex monographs include the FDA pregnancy risk category and/or the simplified Micromedex Fetal Risk Classification. For evaluation purposes we normalized the risk categories of all three sets to the three Micromedex® categories: Micromedex Fetal risk is minimal equivalent to FDA categories A and B, Micromedex Fetal risk cannot be ruled out equivalent to FDA category C, and Micromedex Fetal risk has been demonstrated equivalent to FDA categories D and X.

We calculated accuracy, precision, recall and F1 measure to compare how well the known risk categories in the training set of the FDA-approved SPL documents agree with Micromedex manually curated expert knowledgebase.
To evaluate the performance of the classifier on the test set, we calculated the same measures, using Micromedex as reference standard. We also established a baseline classifier by assigning the majority class label (Fetal risk cannot be ruled out) to the testing set. We tested the statistical significance of the differences in performance measures between the model and the baseline taking into account sample size and the number of classes according to the description by Combrissona 16.

Error Analysis

To analyze the errors of the model we analyzed the differences in classification between the Micromedex set and the testing set for the most extreme case of disagreement, for which Micromedex categorizes the drug as demonstrated fetal risk, and the testing set as minimal fetal risk.

Results

The prescription collection of product labels included 18,341 documents. We excluded 81 labels with a multiple risk category mentions. The distribution of the risk categories for the training set was Class A 1%, Class B 24%, Class C 59%, Class D 11%, and Class X 4%; the training set accounts for 83% of the documents, and the testing set for 17% (Table 3). The distribution for the class labels assigned to the testing set by the classifier was: Class A 2.11%, Class B 15.10, Class C 54.98%, Class D 19.64%, and Class X 8.17% (Table 3)

Table 3: Prescription Drugs: Distribution of Class Labels for Pregnancy Drug Categories.

<table>
<thead>
<tr>
<th>Training Set labels (Gold Standard)</th>
<th>Test Set labels</th>
<th>Correctly assigned by classifier</th>
<th>Misclassified as class (number of labels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold Standard</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A 198(1%)</td>
<td>64(2.11%)</td>
<td>61</td>
<td>C (3)</td>
</tr>
<tr>
<td>B 3,587(24%)</td>
<td>458(15.10%)</td>
<td>370</td>
<td>C(88)</td>
</tr>
<tr>
<td>C 9,050(59%)</td>
<td>1668(54.98%)</td>
<td>1665</td>
<td>B(3)</td>
</tr>
<tr>
<td>D 1,704(11%)</td>
<td>596(19.64%)</td>
<td>579</td>
<td>B(3), C(12), X(2)</td>
</tr>
<tr>
<td>X 682(4%)</td>
<td>248(8.17%)</td>
<td>233</td>
<td>B(3), C(12)</td>
</tr>
<tr>
<td>Total 15,221(83% of total)</td>
<td>3,034(17% of total)</td>
<td>2,908(95.84%)</td>
<td>126(4.33%)</td>
</tr>
</tbody>
</table>

As mentioned above, we selected the sections most likely to contain information about pregnancy risk using ROC. The highest ROC weighted average for the cross validation of the SMO on the training set was for the “Precautions section” (0.99), values for the ROC weighted average for the other document sections are in Table 4.

Table 4: Results of Sequential Minimal Optimization ROC Weighted Average on Document Sections of the Testing Set with 10 Fold Cross Validation

<table>
<thead>
<tr>
<th>Section</th>
<th>ROC Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>How supplied section</td>
<td>0.74</td>
</tr>
<tr>
<td>Indications and usage section</td>
<td>0.89</td>
</tr>
<tr>
<td>Information for patients section</td>
<td>0.90</td>
</tr>
<tr>
<td>Description section</td>
<td>0.90</td>
</tr>
<tr>
<td>Contraindications section</td>
<td>0.90</td>
</tr>
<tr>
<td>Adverse reactions section</td>
<td>0.91</td>
</tr>
<tr>
<td>Overdosage section</td>
<td>0.92</td>
</tr>
<tr>
<td>Dosage and administration section</td>
<td>0.93</td>
</tr>
<tr>
<td>Clinical pharmacology section</td>
<td>0.94</td>
</tr>
</tbody>
</table>
The 15,221 labels in the training set mapped to 685 distinct single ingredients (IN). The 3,039 labels in testing set mapped to 286 single IN. In the training set, we found 37 single ingredients with more than one risk category. Unlike the 81 documents we initially identified as having two risk categories in the text and removed, these 37 ingredients had different categories assigned in different documents, indicating inconsistency in the content labeling across different manufactures. We analyzed the content of two of the 37 ingredients for which we encountered more than one pregnancy risk category in the training set (Table 5).

**Table 5: Ingredients with mention of more than one category in the training set**

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Document Counts for Different Risk Categories for the same Ingredient</th>
</tr>
</thead>
<tbody>
<tr>
<td>RxCui</td>
<td>Ingredient Name</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>2582</td>
<td>Clindamycin</td>
</tr>
<tr>
<td>3992</td>
<td>Epinephrine</td>
</tr>
<tr>
<td>3355</td>
<td>Diclofenac</td>
</tr>
<tr>
<td>4815</td>
<td>Glyburide</td>
</tr>
<tr>
<td>4053</td>
<td>Erythromycin</td>
</tr>
<tr>
<td>7299</td>
<td>Neomycin</td>
</tr>
<tr>
<td>142438</td>
<td>Gentamicin Sulfate (USP)</td>
</tr>
<tr>
<td>10755</td>
<td>Tretinoin</td>
</tr>
<tr>
<td>8134</td>
<td>Phenobarbital</td>
</tr>
<tr>
<td>11002</td>
<td>Urea</td>
</tr>
<tr>
<td>11124</td>
<td>Vancomycin</td>
</tr>
<tr>
<td>10627</td>
<td>Tobramycin</td>
</tr>
<tr>
<td>7213</td>
<td>Ipratropium</td>
</tr>
<tr>
<td>4450</td>
<td>Fluconazole</td>
</tr>
<tr>
<td>6468</td>
<td>Loperamide</td>
</tr>
<tr>
<td>6703</td>
<td>Megestrol</td>
</tr>
<tr>
<td>1223</td>
<td>Atropine</td>
</tr>
<tr>
<td>5487</td>
<td>Hydrochlorothiazide</td>
</tr>
<tr>
<td>33272</td>
<td>phenindimetrazine</td>
</tr>
<tr>
<td>2623</td>
<td>Clotrimazole</td>
</tr>
<tr>
<td>11295</td>
<td>Water</td>
</tr>
<tr>
<td>25789</td>
<td>glimepiride</td>
</tr>
<tr>
<td>7242</td>
<td>Naloxone</td>
</tr>
<tr>
<td>3498</td>
<td>Diphenhydramine</td>
</tr>
<tr>
<td>2409</td>
<td>Chlorthalidone</td>
</tr>
<tr>
<td>1886</td>
<td>Caffeine</td>
</tr>
<tr>
<td>6585</td>
<td>Magnesium Sulfate</td>
</tr>
<tr>
<td>6470</td>
<td>Lorazepam</td>
</tr>
<tr>
<td>10368</td>
<td>Terbutaline</td>
</tr>
<tr>
<td>6628</td>
<td>Mannitol</td>
</tr>
<tr>
<td>6694</td>
<td>Mefloquine</td>
</tr>
</tbody>
</table>
We manually reviewed the Micromedex knowledgebase monographs for 685 ingredients in the training set and 286 ingredients in the testing set.

The SPL training set and the Micromedex reviewed ingredients set had 200 ingredient-class pairs in common, and the Micromedex and the testing set had 238 ingredient-class pairs in common.

The performance measures between the SPL training set and the Micromedex set resulted in: Accuracy 0.94, Precision 0.90, Recall 0.88 and F1 measure of 0.89. The performance of the classifier tested against the Micromedex knowledge base were Accuracy 0.79, Precision 0.66, Recall 0.64 and F measure 0.65. Measures for assigning the most frequent class were Accuracy 0.58, Precision 0.58, Recall 1, and F measure 0.74. The statistical significance for a sample size of 200 and 3 classes exceeds p<0.001.

Error Analysis
We found four drugs in extreme disagreement between the testing set and Micromedex where the drug is clearly contraindicated in Micromedex and the classifier assigned minimal fetal risk. Diazepam: The SPL warning for this drug is in the Warnings Section, and not in the Precautions section used to train the classifier. Both the SPL and Micromedex clearly classify it as contraindicated. Estradiol: The pregnancy section for this drug in SPL refers to no apparent increased risk of birth defects in women who have used the drug in low dose form as contraceptive during early pregnancy, there is no further statement for use of the drug in other stages of pregnancy or at higher doses. Levonorgestrel: The SPL for this drug has no mention of use during pregnancy. In Micromedex there is a clear statement of fetal risks, which affirms that the drug is contraindicated in women who are or may become pregnant. Meprobamate: Both the SPL and Micromedex affirm there is positive evidence of fetal risk but it may be used if the drug is needed in life-threatening situations for the pregnant woman, and there is no other drug that can be effective.

Discussion and Conclusions
We demonstrated that it is possible to automatically classify drug documents into pregnancy risk categories using standard document classification machine learning algorithms, and thus extracting valuable information from free text documents. Several authors have used the FDA SPL documents to extract valuable clinical information using different approaches. Fung and colleagues demonstrated the feasibility of extracting drug indication information from FDA SPL using publicly available natural language processing tools. Futhermore, Khare describes a method to extract structured and normalized indications from FDA drug labels. Culbertson and colleagues used semantic natural language processing (SemRep) to extract Adverse Drug Event Information from Black Box Warnings. Deleger et al. used a hybrid Natural Language processing method to extract indications, contraindications, overdosage, and adverse reactions from FDA SPL documents. To the best of our knowledge, our work is the first to attempt automatic drug document classification based on pregnancy risk categories.

Our work also demonstrates that the automatic classification method provides better precision and accuracy than assuming the most frequent medium-level risk for all drugs with unknown pregnancy risks.

The performance of the classifier was somewhat affected by inconsistencies in the content of the labels for the same ingredients across different manufactures, and by inconsistencies of the document section in which the pregnancy warning statement is included. Inconsistencies in the information in the SPL documents were demonstrated by Duke and colleagues who reported 68% discrepancy in the labeling across manufactures in the Warnings section of bioequivalent medications (identical active ingredient). Duke et al. also report discrepancies in the reported adverse events, post-marketing reports, and even indications differences.

It is not surprising that there are disagreements in the pregnancy risk categorization among similar drugs in the same source. In (Table 5) we included all the ingredients for which more than one risk category is mentioned for the same

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Class</th>
<th>SPL</th>
<th>Micromedex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoproterenol</td>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Methoxsalen</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Budesonide</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Somatropin</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Methylene blue</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Chorionic Gonadotropin</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

We concluded that it is possible to automatically classify drug documents into pregnancy risk categories using standard document classification machine learning algorithms, and thus extracting valuable information from free text documents.
ingredient in different documents, and analyzed in detail two of those drugs. For RxCUI: 6585 RxNormName: Magnesium Sulfate the ingredient is in 15 SPL documents, all of which indicate that the substance is safe for use during pregnancy and is a drug of choice in the treatment of Pregnancy Induced Hypertension, and prevention of seizures in these patients. Only 6 documents, however, label it as category A, 2 documents label it as category C, warning that the use of the drug for more than 7 days can cause bone anomalies in the fetus, and if used 2 hours before delivery the fetus may suffer from severe hypocalcemia. Further, 7 documents label the same substance as category D based on the animal studies showing that the prolonged use of the drug can cause bone alterations and alter reproductive capacity of the fetus. Another example of multiple drug categories for the same ingredient is RxCUI: 4450 RxNormName: Fluconazole, for this drug tablets for vaginal application have a risk category A, while tablets for oral treatment have a risk category C. The drug is not expected to be absorbed systemically when used in topical or vaginal applications, therefore it is not expected to cause the teratogenic effects demonstrated in animals with oral treatment, but there is no conclusive knowledge on the amount of drug absorbed systemically with vaginal application.

Other examples of disagreements in the reference standards are the assignments of a single risk category to the drugs with different risks to the developing fetus depending on the gestational age. We excluded from our study documents that included more than one category, but some manufacturers label the drug with only the highest risk category as we demonstrated with the analysis of the text for the 15 labels containing Magnesium Sulfate. This analysis indicated that in the future, gestational age needs to be taken into consideration, when automatically assigning pregnancy risk categories to drugs.

Federal Register Content and Format of Labeling for Human Prescription Drug and Biological Products; Requirements for Pregnancy and Lactation Labeling 22 issued on December 4, 2014 removes the pregnancy risk category labeling from the Standard Product Labels and implements the requirement to include patient registry information when available. However, the FDA and its equivalent Australian organization have attempted this change for several years. Progress is slow and not likely to happen without the aid of automatic methods. Patient registries on prescription medication use are slowly growing; the information obtained from these data sources will require curation before it is applicable to clinical documentation. Patient registries will also benefit from automatic classification systems similar to the one tested by us. Moreover, patient registries are focused on drug classes and the information is not yet available to the public.

Our study demonstrates that it is possible to extract useful clinical information from free text using automatic machine learning algorithms, although more work is needed, as indicated by the 89% F1 score when comparing the known SPL category assignments and Micromedex, which could be considered an achievable agreement for machine learning methods. Additional experiments are needed to test if the accuracy, precision and recall can be improved by deeper understanding of the context, such as the gestational age, and through additional syntactic and semantic features of the SPL precautions sections. Automatic classification of drug labels lacking pregnancy risk category information can aid clinicians and patients who are attempting to identify the effects of drugs similar to those for which the information is available in text format. Our approach could also be of general interest for extraction of other information that might be incomplete in some drug labels, such as contraindications or side-effects, e.g. secondary effects of Tricyclic antidepressants (TCAs) depend on pharmacological action on histamine or serotonin receptors, an automatic system that could automatically classify these drugs could be of great benefit at the point of care. Information on drugs changes every day, continued research both in the pharmaceutical industry and in clinical practice unveil previously unknown information, the FDA is continuously releasing drug safety communications 23. It is important to provide a mechanism by which identical drugs can be automatically classified utilizing the existing knowledge.

Limitations
Our work is limited to prescription drugs, however, over the counter drugs and homeopathic medications have potential to cause harm on the developing fetus as well and need to be explored in the future.

Acknowledgments
Institutional: This work was funded by the Intramural Research Program of the National Library of Medicine.
Individuals: Phil Wolf BS
References


Electronic Health Records (EHRs) have increased the utility and portability of health information by storing it in structured formats. However, EHRs separate this structured data from the rich, free-text descriptions of clinical notes. The ultimate objective of our research is to develop an interactive progress note that unifies entry, access, and retrieval of structured and unstructured health information. In this study we present the design and subsequent testing with eight clinicians of a core element of this envisioned note: free-text order entry. Clinicians saw this new order-entry paradigm as a way to save time and preserve data quality by reducing double-documentation. However, they wanted the prototype to recognize more diverse types of shorthand and apply default values to fields that remain fairly constant across orders, such as number of refills and pickup location. Future work will test more complex orders, such as cascading orders, with a broader range of clinicians.

Introduction

Electronic Health Records (EHRs) have greatly increased the utility and portability of health information by storing it in structured formats. There are numerous benefits to having patient records in these computer-readable formats including being able to perform targeted searches, exchange data between institutions, and provide real-time clinical decision support. However, current EHRs rely on an aging paradigm of windows, tables, and menus to collect and display this information. Since these elements take significant time and attention to use, providers spend precious moments navigating EHRs when they could be engaging with patients or documenting visits.

Moreover, as health records have moved from paper to computers, the progress note has lost centrality as an instrument of care. Patient information is now scattered across numerous parts of the EHR, strongly divided into structured and unstructured sections. This fragmentation imposes a range of new constraints on physicians’ documentation and is a key cause of errors. It can also lead to duplicate work, for instance when an order placed via a drop-down menu must later be manually documented in a note.

The ultimate goal of our work is to develop a novel EHR paradigm around interactive notes that unify entry, access, and retrieval of structured and unstructured health information. This study tests one feature of this note-centric vision: free-text order entry. The particular research questions we sought to answer with this study were: (i) what types of shorthand do clinicians use when placing free-text medication orders? (ii) what types of functionality do they expect from computerized free-text order entry? and (iii) do they see free-text order entry as a useful addition to their note?

Background

Moving medical records into a computational medium has conferred numerous benefits on the medical community including remote access to patient records and real-time clinical decision support. EHR use has also been associated with higher quality of care. However, there has long been recognition that by separating information into discrete categories displayed on separate pages, EHRs hinder clinicians from making inferences across categories, such as when referencing lab results to determine why an order was placed.

Despite this fragmentation, clinical notes remain a central component of EHRs. They unify the record by describing a patient’s history, interpreting test results, and justifying care plans. They are also where clinicians spend much of their time. Our prior research on outpatient visits across several Veterans Affairs hospitals found that clinicians spent the majority (58%) of their EHR time in notes. Despite being central to clinical care, notes are increasingly
difficult to write as they need to include large amounts of structured data, already documented elsewhere in the record, to meet regulatory and billing requirements and defend against lawsuits\(^6\).

Researchers in biomedical informatics and human-computer interaction have sought to reduce this duplicate work by making it easier to carry information across notes\(^7\) or copy structured information into a note\(^8\). Moreover, some EHR vendors have started to provide means for quickly importing information such as vital signs using in-note commands, such as Epic’s Smart Phrases\(^9\). This prior work has focused on pulling structured information into the progress note, but little attention has been given to the reverse conversion of creating structured data in real-time from note text. This concept is intimately connected to previous work on information retrieval, so much so that after testing the in-note information retrieval system in\(^8\), one participant remarked that they would “like to place orders for medications and tests” from within the note. The Regenstrief Institute has begun to explore this direction with Medical Gopher, an EHR that recognizes medication keywords written in a note and suggests placing related orders\(^10\). However the Medical Gopher does not recognize order details such as dosage or schedule and clicking on the suggested order takes users to a separate screen with menus to specify these details.

One of our objectives in creating an interactive progress note is to let clinicians take actions, such as placing an order, as they are creating their note. Towards that objective, we developed and tested a prototype note editor to observe how clinicians would expect to place medication orders using free-text. This research follows iterative prototyping methods common in human-computer interaction in which early prototypes are tested with potential end users to identify usability issues and validate that the design addresses the identified need\(^11\).

**ActiveNotes Prototype**

To test clinicians’ methods of free-text order entry, we developed ActiveNotes (Figure 1). At its core, ActiveNotes is a note editor that lets clinicians create and edit notes. Like many EHR note editors, there are no facilities for rich-text formatting such as bolding or changing font size.

![Figure 1. The ActiveNotes Prototype in use. The center white area is the note editor. When a user types “#med”, a medication ordering dialog appears, allowing free-text orders to be placed. Text typed in this dialog is automatically mapped to structured information as indicated by the highlighted components above the medication order (i.e., ‘Med’, ‘Strength’, ‘Schedule’, ‘Days’, and ‘Refills’ in the example). The patient information to the left of the note is static and does not reflect note content.](image-url)
ActiveNotes’ core strength is in its ability to parse semi-structured text into full medication orders. To start an order, users type the tag “#med” which opens a small order specification dialog on top of the note (Figure 2). This window provides a search bar where clinicians can type their order. ActiveNotes expects the medication and dose information to be entered first and provides auto-completion for these fields based on an underlying medication database built around RxNorm and NDF\textsuperscript{12}. As text is entered in the search field, it also appears in the note wherever the #med tag was invoked. After entering the medication and dose fields, the order is complete enough for checkout and clinicians can hit their “Enter” key to finish the order. Alternatively, they can specify other aspects of the order such as form, route, schedule, refills, etc. As each part of the order is recognized, the dialog highlights the components’ corresponding label in blue in the popover window. Users can also access a standard order specification drop-down menu by clicking the triangle to the right of the search box, though they are not encouraged to use this feature.

![Figure 2. Inline order-entry. (a) Autocompletion of a medication name (b) When ActiveNotes recognizes a field, it highlights a corresponding label above the search box in blue (c) drop-down ordering](image)

After a medication is ordered through the dialog, it can be edited or deleted through an action menu that appears when users right-click or hover over the order text (Figure 3). Anticipating future uses of ActiveNotes, we have developed order specific tasks including refill, reorder, and discontinue that can all be accomplished in a single click. These features were not tested in this study.

![Figure 3. Orders can be modified by hovering over or right-clicking on the order text](image)
Once finished with their note, users can go to the “Order Checkout” page to review and sign any orders they placed (Figure 4). Here, ActiveNotes asks for any additional information needed to fully specify the order, but it also fills in default values for some fields that were not specified in the note, such as pickup location or number of refills. ActiveNotes currently supports only basic medication orders but we plan to expand it in the future to support more complex medication orders, such as weight-based orders, as well as orders for lab tests, radiology, and consultations.

Figure 4. ActiveNotes’ Checkout screen where orders can be reviewed and edited before being signed. The red box highlights missing information. Users can either enter it here, or go back to the note to complete the order.

ActiveNotes is build on top of standard web technology stack and medical ontologies (Figure 5). The user interface was developed in HTML5/CSS/JavaScript and exploits open source libraries to enable more complex interactions and UI elements. The middle layer exploits standard servers and frameworks such as SQL, REST and PHP to recognize the order fields and provide autocomplete suggestions in real-time.

Figure 5. ActiveNotes is build on standard web technologies and medical ontologies.
Evaluation

We tested ActiveNotes through a formative usability test with eight clinicians from San Diego’s Veterans Affairs hospital and UC San Diego Medical Center. The clinicians’ specializations included internal medicine (3), infectious diseases (4), and surgery (1). Our participant pool included both inpatient and outpatient practitioners, with two fellows and six senior clinicians. Each test lasted 30 minutes or less. Clinicians were first shown a short video explaining ActiveNotes’ features. Following the task scenario method of usability testing, they were then given two short descriptions of canonical outpatient visits and asked to create a progress note for each visit in ActiveNotes. Each scenario gave basic demographic and assessment information and asked the clinicians to place four new medication orders from within their note. The orders were largely phrased without standard shorthand (e.g. PO, BID) so we could observe the shorthand clinicians naturally used. After completing both progress notes, the clinicians were asked to comment on their experience with ActiveNotes. We tracked how long it took clinicians to successfully place each order, the shorthand they used while placing each order, any unsupported uses of the system they attempted, and their comments on desired features and overall usefulness of ActiveNotes.

Results

Time to Order

Clinicians took, on average, 59 seconds to place each order, though this time varied greatly between clinicians and orders. For example, P3 averaged just 30 seconds to place each order whereas P8 averaged 1 minute 56 seconds. Also, it took clinicians an average of 39 seconds to place an order for “Simvastatin 20mg PO at bedtime daily. 30 tablets to be dispensed with 3 refills” whereas it took 1 minute 17 seconds to place one for “Aspirin 81mg once daily. 100 tablets to be dispensed with 3 refills”. The Aspirin order took longer because ActiveNotes did not recognize the “asa” as shorthand for “Asprin”, which delayed several clinicians.

Broadly, time to order depended on whether the clinician initially used shorthand that ActiveNotes did not recognize as pertaining to a particular field. For example, whereas ActiveNotes could parse “#30” as quantity information, it could not parse “Dispense: 30”. The orders that took the longest to complete were those in which clinicians tried to use several different shorthands that ActiveNotes did not recognize and then resorted to using its fall-back dropdown menu (accessible by clicking the triangle to the right of the search box). Since one of our main objectives was to identify the shorthand clinicians naturally use, we did not instruct them on what type of shorthand ActiveNotes was programmed to recognize and let them decide when they wanted to fall-back on the dropdown menu.

Shorthand

Clinicians used a variety of shorthand when placing free-text orders. Representative variants are shown in Table 1. Whereas the medication, dose, form, route and days fields saw little variation, schedule, quantity, refills, and pickup information was entered in a number of different ways.

Table 1: Shorthand used when placing orders

<table>
<thead>
<tr>
<th>Information</th>
<th>Shorthand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medication</td>
<td>Aspirin, asa (abbreviation)</td>
</tr>
<tr>
<td>Dose</td>
<td>100mg, 75mcg</td>
</tr>
<tr>
<td>Form</td>
<td>tab, enteric coated</td>
</tr>
<tr>
<td>Route</td>
<td>po, oral</td>
</tr>
<tr>
<td>Schedule</td>
<td>bid, at bedtime, once a day, 1x/day, prn for dizziness</td>
</tr>
<tr>
<td>Days</td>
<td>30 days</td>
</tr>
</tbody>
</table>
Functionality

As is common with prototypes, our participants wanted ActiveNotes to do more. We particularly value feedback in this context since the expected functionality can guide and inform future development. We uncovered six recurring function requests; the first three regarded time saving assistance, the last three address broader integration with the EHR. The clinicians volunteered these comments and were not asked about them specifically.

Default Values - Five clinicians wanted ActiveNotes to have default values for fields like schedule, refills, and pickup. While ActiveNotes’ refill value defaulted to “3 refills” and pickup value to “Window”, it did not have a default value for schedule (such as AM). Furthermore, there is no way to see when placing an order in the note that ActiveNotes will fill in these default values on the Checkout screen unless otherwise specified.

Auto-populated Values - Four clinicians wanted ActiveNotes to automatically populate form and quantity information whenever possible. For example, ActiveNotes should automatically mark ‘Tab’ if the specified medication and dose can only come in a tablet form. Also, after entering a schedule and duration of “BID for 30 days”, ActiveNotes should assign a quantity of 60.

Order Entry Invocation - Four clinicians wanted a simpler way to invoke the order entry window than the “#med” syntax. Clinicians either wanted to type out the full order and then hit a hotkey to tell ActiveNotes to parse the preceding text, or only wanted to invoke the “#med” once per note and fill out multiple orders in a row.

Information Retrieval - Two clinicians wanted ActiveNotes to support rich information retrieval such as “retrieve most recent colonoscopy”.

Additional Orders - Three clinicians expected ActiveNotes to handle additional order types including radiology, labs, and consults.

Additional Parsing - One clinician wanted ActiveNotes to be able to parse other note text, such as active problems, and save it to the appropriate part of the EHR.

Usefulness

Clinicians saw the current implementation of ActiveNotes as useful in two distinct ways. First, they thought it was useful to not have to switch between sections of the EHR to document and place orders. As one clinician remarked:

“I love not having to go out of the screen [to place orders]” - P1

Secondly, clinicians saw directly linking orders to documentation as a way to avoid overlooking orders:

“The thing that I really like about this is... having the order directly tied to the documentation of that order in the note. The issue I run into sometimes is that I write my note and I'm waiting to do my orders and then I have to… make sure... all the things I said I was going to do in the note I actually order.” - P3

More broadly clinicians saw value in using interactive notes to populate the EHR with structured information:

“If this had the ability to take everything in the note and just automatically download it... if you just did like #activeproblems from the note it would just put it all into that section... that would dramatically improve efficiency because there's a lot of that duplication that we're currently doing.” - P7

This is opposed to the current model of importing structured information into the note:
As it currently stands, it's the opposite. You have to first put in everything in the active problem list or medication list, and then you can populate it into the note by using smart keystrokes, but it would be nice when you're initially seeing someone to not have to write the whole note and then repeat everything” - P7

Discussion

Our first research question asked what types of shorthand clinicians use while placing free-text medication orders. Whereas the shorthand for some fields was standardized (e.g. ‘asa’ for aspirin, ‘mg’ for milligrams) fields such as schedule and number of refills were specified in a number of different ways. From this observation, we realize that it will be important for interactive notes to embrace a rich set of terminology that goes beyond current ontologies, such as RxNorm, to include terms for schedule, refills, and pickup. As a first step, we propose iteratively testing ActiveNotes through large-scale online deployments to capture a broad range of shorthands.

Our second research question asked what types of functionality clinicians expect from a computerized free-text order entry system. The six recurring functionality requests fell into two categories: time saving assistance and broader integration with the EHR. Along the lines of time saving assistance, clinicians wanted to specify their order with as little typing as possible. This principle can be seen in their desire to invoke the ‘#med’ dialog only once and place multiple orders in a row. It can also be seen in their desire for ActiveNotes to recognize when a particular dose of a drug only comes in one form, or to automatically calculate quantity given a prescription’s schedule and duration. It is worth noting that those fields that clinicians wanted the most assistance with were also those that exhibited the most varied shorthand. Clinicians also wanted ActiveNotes to be more broadly integrated with the EHR. This included both extending the types of orders it recognized to include labs, imaging, and consults and extending ActiveNotes’ parsing to cover other types of structured information such as active problems and family history.

Our final question asked if clinicians saw free-text order entry as a useful addition to their note editor. They saw free-text order entry as being useful in a number of ways including being a potential time saver, less distracting than form-based input, and requiring less navigation compared to current EHRs. Furthermore, they thought it could save them from needing to enter information twice, once in a structured format, and then in an unstructured format. Textual entry may also require less attention than form-based or drop-down based entry, letting providers focus more on their patients if they choose to document during the patient encounter.

There are also potential technical benefits to letting clinicians tag note content as structured data compared with post-hoc Natural Language Processing (NLP). First, it enables the parser, whether it is looking for orders or a problem list, to use a more targeted ontology for recognizing terms. Second, it adds structure to the note itself by marking which parts of the note refer to medication, conditions, and so on. Such tagging could assist with later NLP or the development of richer note interactions related to targeted search, filtering, and highlighting. Thirdly, order entry dialogs could be expanded to include real-time clinical-decision support, potentially catching errors and drug-drug interactions while notes are being written on rounds, not while orders are being placed later.

Looking towards future designs, we observed that free-text entry of structured information is not a familiar interaction paradigm as compared to menu and drop-down driven user interfaces. While our subjects were not constrained in what they typed, some had difficulty understanding the full functionality of the system. One participant did not grasp that ActiveNotes would let him document and place orders at the same time. Instead, he wrote each order in the note twice, first with fairly standard shorthand (e.g. po, #30) while “documenting” and then with less standard shorthand when “ordering” (e.g. Dispense: 30 ). Beyond developing a robust interactive note, it will take time and experience before some clinicians are comfortable with this new note-creation paradigm.

Pilot tests before this study also revealed that for free-text entry to work, careful thought has to be given to its exact mechanics. For example, which keys users press to complete an order, how they select an auto-completed phrase, and how the system shows that it recognizes part of the order all need to fit clinicians’ expectations.

Conclusion

This study takes a step towards our ultimate objective of developing an interactive progress note. We tested one critical feature of such a note, free-text order entry, and found that (i) clinicians use a variety of shorthand when placing free-text orders, particularly when specifying schedule, pickup location, and number of refills, (ii) clinicians
want to specify a minimum number of fields and rely on default or auto-populated values to fill in routine information, and (iii) clinicians see free-text order entry as a useful addition to their progress notes with the potential to save time and ensure orders are not overlooked.

In the future, we plan to test ActiveNotes with a broader range of clinicians across locations and disciplines, letting us observe a wide range of shorthands so we can develop a more robust order parser. Since ActiveNotes is built on web technology, we can easily deploy it on a larger scale and track visitors’ interactions with the site. Secondly, we plan to extend ActiveNotes’ ordering capabilities to include more complex medication orders, such as weight-based orders and cascading orders, as well as consults, labs, and imaging. Finally, we plan to test ActiveNotes on mobile platforms, exploring the usage of dictation for text input.

We posited that introducing a new interaction technique based on within-note free-text data entry could help solve a variety of problems that EHR users face. To investigate this claim, we designed and tested a novel note-centric approach to medication order entry and found that while work remains before being deployed in a live EHR, our in-line free-text approach was validated by representative end-users. By considering the note as the central element of the EHR and incorporating interactions and operations that typically span multiple parts of EHRs, we have an opportunity to transform documentation from being a complicated and time-consuming clerical task to a more natural interaction with the EHR that is closer to both the clinician and the patient, supporting increased efficiency, fewer errors, and ultimately, better health outcomes.

Acknowledgements

This research was supported by VA Transformational Initiative #16 and the National Library of Medicine training grant T15LM011271. We thank our volunteer physicians for their valuable feedback.

References

Improving Continuity of Care via the Discharge Summary

Farrant H. Sakaguchi, MD, MS1, Leslie A. Lenert, MS, MD2
1Intermountain Healthcare, Salt Lake City, UT; 2Medical University of South Carolina, Charleston, SC

Abstract

Discharge summaries (DCS) frequently fail to improve the continuity of care. A chart review of 188 DCS was performed to identify specific components that could be improved through health information technology. Medication reconciliations were analyzed for completeness and for medical reasoning. Documentation of pending results and follow-up details were analyzed. Patient preferences, patient goals, and the handover tone were noted. Patients were discharged on an average of 9.8 medications, only 3% of medication reconciliations were complete and medical reasoning was frequently absent. There were 358 pending results in 188 hospital discharges though only 14% were mentioned in the DCS. Documentation of clear, timely follow-up was present for less than 50% of patients. Patient preferences, patient goals, and lessons learned were rarely included. A handover tone was in only 17% of the DCS. Evaluating the DCS as a clinical handover is novel but information for safe handovers is frequently missing.

Introduction

Discharge summaries (DCS) can decrease readmissions and adverse events. When a DCS was available at follow-up after hospital discharge, van Walraven et al. found a trend for decreased readmissions (relative risk 0.74), and when it was unavailable, Li et al. found a statistically significant 79% increase in the readmission rate within 7 days.1,2 The incidence and severity of adverse events (AEs) after hospital discharge were found to be preventable or ameliorable with better communication from the hospital to community providers by Forster et al.3,4

However, DCS frequently fail to reach their potential to improve the continuity of care. A complete list of medications and discharge diagnoses are among the most important components of the DCS.5–9 Yet discharge medication discrepancies are common, even when a pharmacist was involved.10 Inaccuracies in discharge medication lists, medication changes, and reason for medication changes were found in only 36%, 30%, and 38% of DCSs by Legault et al.11

The exclusion of tests pending at the time of discharge is another weakness of DCSs. Were et al. found only 13% of DCS reported all pending tests and that only 16% of all pending results were reported.12 Roy et al. found results returning after discharge for 41% of patients, 43% of which were abnormal, two-thirds of which were potentially actionable.13

Moore et al. found the diagnostic and follow-up plans in the inpatient medical record were not completed by the outpatient physicians 41% of the time.14 However, the recommended workups were documented in only 46% of the DCS.15

van Walraven et al. also found that DCSs were not reaching follow-up physicians at 85% of the follow-up visits.16 Two thirds of those cases were because the DCS was simply never sent to the outpatient physician. In only 59% of cases studied by Were et al. was there sufficient information to ensure delivery of the DCS and results.12 Though there is no clear standard in the literature, the timeliness of the delivery of the DCS is frequently related to its quality.5,6,17–19

As the lack of timely postdischarge follow-up has been associated with hospital readmissions and emergency department visits, the Society for Hospital Medicine has endorsed the scheduling of follow-up appointments and including the details in the DCS.9,20

Multiple metrics have been developed to measure the quality of the DCS. van Walraven and Rokosh surveyed 100 physicians who identified admission and discharge diagnoses, discharge medications, follow-up plan, pending labs,
1-2 pages total length, receipt within four days, active medical problems, and outstanding social issues as being components of a high quality DCS. Horwitz et al. scored the timeliness of completion separately from the transmission of the DCS. Rao et al. focused on the quality of the communication (the clarity of the DCS, exclusion of irrelevant material, and consistency of the documentation). Stetson et al. developed the nine-question physician documentation quality instrument (PDQI), but still felt unable to describe an ideal DCS.

However, a comprehensive framework for DCS that emphasizes clear communication to support continuity of care is missing from the literature. Kripalani et al. specifically identifies communication deficits between inpatient and outpatient physicians as being common and with the potential to adversely affect care. Easily coded data such as accurate medication lists and tests pending at discharge are essential. Yet a comprehensive data dump should not obscure the context about patients; the medical reasoning behind the selection or duration of a treatment regimen or recommended next steps for the workup are also critical. Prior metrics have been useful to recognize specific information components, but without putting them in the context of a handover; in a handover both the responsibility and the information needs to be passed from the inpatient provider to the outpatient provider. Patient handovers involve 1) the sharing of information, 2) exchanging responsibility for a patient, and 3) transferring an understanding of the medical decision-making. As Moore et al. found, the outpatient provider is unable to complete recommended workups if those plans are not communicated; continuity of care is more feasible when there are effective handovers.

With the growing use of EHRs, it is important to consider how to leverage health information technology to ensure efficient communication between clinicians. Administrative and encoded clinical data should be automatically inserted into documentation to ensure the consistency of useful documentation. But such boilerplates should not come at the cost of clear clinical communication or by excessively long documents. Thoughtful summarization may emphasize the gist of the hospitalization and follow-up needs efficiently and effectively in a concise, cohesive narrative. Since DCS may be the only communication between the hospitalist and the primary care provider, they should be treated as handovers.

The aim of this study is to characterize the data in DCS through a retrospective chart review for its availability and its effectiveness of facilitating continuity of care. We focused on the medication list in the DCS, pending results, coordination of follow-up, and evidence of a clinical handover.

Methods

Setting: Data for this study come from an urban tertiary academic referral center in the intermountain west. Most patients are English-speakers. Patients admitted to the general medicine services were treated by four teaching or one attending hospitalist services. Each of the four teaching teams consisted of an attending, a senior resident, two interns, and two medical students. The attending service consisted of an attending and a senior resident. The housestaff generally dictated the DCS though typing was allowed.

Participants and Data Sources: To stratify the experience-level of the housestaff, three convenience samples were defined. The first 122 consecutive patient discharges were from the general medicine teaching service starting August 1, 2010, the next 33 consecutive discharges started May 1, 2011 from the same service, and the next 33 consecutive discharges starting May 1, 2011 from the attending service; i.e. the DCSs were created by new interns, experienced interns, and senior residents, respectively.

Discharge summaries, computerized-provider-order-entry (CPOE) orders, lab results along with date-timestamps of the order and the availability of the result, patient demographic information, length of hospital stay, and diagnostic codes used for the billing record were collected from the enterprise data warehouse (EDW). An adapted Charlson Comorbidity Index (CCI) was calculated from diagnoses and comorbidities in the billing record for the date of discharge.

Chart Abstraction and Data Analysis: Each DCS was abstracted into four main categories: a) the medication list, b) labs pending at the time of discharge, c) posthospitalization follow-up, and d) the clinical handover for continuity of care. The data abstraction was performed by one author (FS), a biomedical informatics postdoctoral fellow and practicing family physician.
The medication list was evaluated for its completeness, and each medication was evaluated for its status and for the presence of any relevant medical reasoning. A “Complete” medication list should indicate the status of each medication being “Continued,” “New,” “Changed,” or “Discontinued.” If the status of a given medication was “Unclear,” it was categorized as such and the medication reconciliation was considered “Incomplete.” The presence of medical reasoning regarding the indication, selection, or duration for each medication was noted. The number of medications per DCS, the proportion of “Complete” medication lists, and the proportions of the status and the presence of medical reasoning for each of the medications are reported.

Labs pending at the time of discharge were measured by treating the DCS and the EDW record of CPOE orders as separate sources. From the DCS, every mention of a pending result (e.g. a pending blood culture) and recommended follow-up test (e.g. follow-up INR or scheduled sleep study) was recorded. From the EDW, every laboratory order with pending results at the time of discharge was identified. The two lists were compared for each patient, item by item. For each pending result, there was one possible condition: a) reported in both the DCS and the EDW, b) reported in the DCS but not the EDW, or c) reported in the EDW but not the DCS.

Hospital follow-up was measured by the identification of a follow-up provider and the appointment. e.g. “Dr. Smith” was considered “General” whereas “Dr. George Smith in Clinic 2” was considered “Specific.” Follow-up appointments were categorized as a) within 14 days, b) beyond 14 days, c) not scheduled but recommended within a specific timeframe, d) not mentioned, or e) deferred due to patient transfer.

We developed a framework to assess “continuity of care” and “handover tone” as these qualitative concepts are missing from other metrics. To assess for a “continuity of care,” we looked for the presence of Patient Values, Preferences, Goals, and “Lessons Learned” in the DCSs. e.g. “a patient values independence more than safety,” “a patient preferred nursing home A due to proximity to family,” “the patient’s goal is to return to her own home before the holidays,” or “the key lesson from this hospitalization is...,” respectively. We drew from Weir’s adaptation of Hollnagel and Wood’s contextual control model (COCOM) to describe components of an effective handover. “Handover Tone” was operationalized as when there were three of the following: a) a cohesive story, b) predictions and guidance for the patient’s clinical trajectory, c) an explicit plan moving forward, d) key parameters to monitor, e) clear medical reasoning, or f) a holistic perspective about the patient. The frequencies of Patient Values, Preferences, Goals, “Lessons Learned,” and a “Handover Tone” in the DCSs are reported.

Statistical analysis was performed to evaluate differences between the cohorts. A 75% probability was estimated that a given categorical measure, such as whether the primary care provider was identified, would be present in the DCS. The minimal sample size to detect a difference between categorical variables is calculated by $N = 4z^2p(1-p)/\Delta^2$. For $z_{0.05} = 1.96$, probability $= 0.75$, and precision $= 0.2$, the necessary sample size to detect a difference between would be 51. The chi-squared test evaluated for statistical difference between the three cohorts for categorical variables (gender distribution, whether the primary care provider was identified, and whether the patient was discharged to home, the length of the DCS, and the time from discharge to dictation of the DCS. An ANOVA was used to evaluate for gross differences between the cohorts regarding patient age, length of stay, and the Charlson Comorbidity Index.

The data were summarized in a fishbone diagram to illustrate quality gaps and informatics opportunities to improve the DCS. Each “bone” of the diagram represents a component of the DCS, that when missing or incomplete, contributes to creating discontinuity of care; the modes of failure do not occur in a simple linear process.

Results

Using ANOVA (for continuous variables) and chi-squared tests (for categorical variables), we found no statistical differences for the age, length of stay, comorbidity index, DCS length in words, time from discharge order to dictation, gender-ratio, identification of PCP, or discharge to home between the three cohorts, as summarized in Table 1. Grouping all discharges together, 95 or 51% of the patients were female, with an average age of 58 (SD 18) years, length of stay of 4 (SD 5.2) days, and a modified Charlson Comorbidity Index of 3.2 (SD 2.5). The mean length and time until dictation of all the discharge summaries were 762 (SD 336) words and 2 (SD 3.9) day, respectively.

Table 2 summarizes the number of medications, their status, and the availability of the related medical reasoning. 188 patients were discharged on an average of 9.8 (SD 5.7) medications. Only five out of 188 discharge summaries
were “Complete,” explicitly reporting the status of every medication. In the 61 discharge summaries when it was explicit, an average of 2.1 medications were discontinued. Seven discharge summaries failed to name any specific medications by either omitting the medication list or indicating “all medications were continued.” An analysis of the differences between the discharge summary and the pharmacist-performed medication reconciliation at discharge is reported elsewhere and is currently under review.

In the 188 discharges, 358 lab results (an average of two results per patient), were pending at the time of discharge from the hospital. 31 (9%) of the pending results were identified in both the discharge summary and by the EDW, 17 (5%) were identified by the discharge summary but not by the EDW, and the remainder, 310 (86%), were identified in the EDW but not in the DCS.

Not only are pending lab results commonly omitted from the DCS, those reported correlate poorly to those identified in the EDW, shown in Table 3. Of the 188 patients, only two (1%) cases had an exact match between the dictated discharge summary and the electronic record. In one of those cases, there were pending results and in the other case, the summary stated correctly there were no pending results.

Another 61 (32%) of cases implicitly matched where no pending results were mentioned in the DCS and none were found in the EDW. In the remaining 125 (66%) of cases, there was some discrepancy between the pending results noted by the discharge summary and reported by the EDW.

31 (16%) patients were discharged on warfarin. In only one third of those discharge summaries (10 patients) was there mention of a follow-up INR or how anticoagulation was to be managed as an outpatient.

A primary care provider (PCP) was identified specifically, generally, or not at all in 34%, 27%, and 39% of DCS, respectively. Since some DCS mention a PCP and specialist, 45% of all DCS identified at least one specific follow-up provider.

The DCS contained scheduled follow-up appointments for 88 (47%) and 11 (6%) of patients within 14 days and beyond 14 days, respectively. For 48 (26%) patients, some follow-up was recommended while no follow-up was mentioned for 29 (15%).

There were references in zero (0% of discharges), 11 (6% of discharges), 2 (1% of cases), and 5 (3% of cases) to patient values, patient preferences, patient goals, and lessons learned, respectively. An example statements of patient preferences was, “It was a good talk with Palliative Care; and the patient is to remain DNR/DNI, but would still like interventions other than that to help keep her healthy.” An example of a patient goal was, “He was transferred to [skilled nursing] for ongoing PT and OT therapy in hopes to regain his strength and eventually return home.” An example of a lesson learned was, “It also should be noted that the patient was only taking three medications at home once daily and it was unclear which medications he was taking and therefore we restarted the above medications and set the patient up with home health for assistance with medication administration.”

Handover tone was also uncommon. While a cohesive story and clear medical reasoning of the hospitalization was available in 151 and 156 cases, respectively, predictions and guidance for the patient’s clinical trajectory, an explicit plan with key parameters to monitor or a clear process to follow, or a holistic perspective of the patient was present in only nine, 22, and 20 of the cases, respectively. A handover tone was achieved in only 32 (17%) cases.

Figure 1 summarizes the modes of failure to provide continuity of care in discharge summaries in the topic areas of Medications, Follow-up, Pending Results, and Clinical Handover. These points were chosen as based on the data available in our study as well as the potential to improve these areas through health information technology. It helps to illustrate that improving discharge summaries will require a socio-technical solution.24

Strengths and Limitations
The key strength of these findings is framing the analysis of the DCS primarily as a handover. Also, the mixed analysis method of performing a chart review and comparing relevant findings to results from an EDW query is relatively unique.
One of the greatest limitations to these findings is the use of a single data abstractor. It was felt that a physician or possibly a nurse was necessary in order to be familiar with the clinical context. Unfortunately, other clinical resources were unavailable for the data abstraction.

Another potential limitation was the attempt to stratify the experience levels of the physicians. Some might argue that a random sampling may be more representative. However, we found no differences between the three cohorts in terms of patient demographics, case complexity, or general analysis of the DCSs themselves.

Some might express concern about the generalizability of the results from a single academic institution or the use of a specific EHR. However, the hospital uses a common, widely implemented EHR and has CPOE fully implemented. Future work may reproduce this analysis at other sites.

Discussion

We found several quality gaps in the DCS at our institution, consistent with the literature. The context for these opportunities to improve are the DCSs created at an institution recognized by the U.S. News & World Report’s Best Hospitals and the University Health System Consortium as among the best in quality. It is also noteworthy that the institution’s readmission rate is below the national average.

Based on the DCS, we conclude that continuity of care was not a consideration at discharge; a specific follow-up provider was identified only 45% of the time, lower than the 67% observed by Were et al. Without identifying a specific follow-up provider, it becomes impractical and a legal liability to forward important information regarding the hospitalization contained in the discharge summary. These conclusions are inline with the low continuity scores observed by Van Walraven et al. Were et al. found 16% of pending tests reported in the discharge summary and Roy et al. found that 41% of patients discharged had pending labs where we found 67%. Were et al. found that discharge summaries reported only 48 (14%) of all pending results but that 17 (5%) were not identified by the EHR. Our conclusions were consistent with those of Walz et al.; the majority of pending test results were microbiology tests with the majority of those specifically being pending cultures.

Despite consensus that anticoagulation at transitions of care should be carefully coordinated, there was little evidence of this in our data with 68% of warfarin patients lacking mention of a follow-up INR.

Several prior efforts to improve discharge summaries have focused on the content of the discharge summaries such as through the use of checklists. A related approach that has shown improvement in the rated quality of discharge summaries is through formal teaching interventions. O’Leary was successful in creating a draft electronic discharge summary template that would automatically insert specific data elements and found an overall improvement in the quality and timeliness of discharge summaries.

A focus on conveying medical reasoning in the discharge summary has been essentially missing altogether from the literature. Several studies seem to recognize this problem, but fail to identify the need to explicitly communicate medical decision-making. The paucity of medical reasoning became evident in our chart review. However there is no “best” place to document the medical reasoning. Such information is not universally applicable; there is no data whether medical decision making fits best as a comment with each individual therapy within individual modules such as CPOE, the medication list, or even the problem list, versus combined sections of a history and physical, progress note, or discharge summary. Even the Moore et al. documentation of failure to communicate loose ends and intended diagnostic plans did not address the concepts of “lessons learned” or the clinical trajectory of a patient, patient preferences, or patient goals. The absence of this global perspective for the patient once he or she left the hospital was evident in our study.

The discharge summary may be the only clinician-to-clinician communication when the patient leaves the hospital and thus appears to be the most practical form for a handover. Our findings suggest that not only are DCSs imperfect as information containers, they are poorly suited as clinical handovers. While the data could be more complete by leveraging the EHR alone, handover tone will likely require a paradigm shift. The real failing of the discharge
summary is not the mere absence of data, but rather the failure to recognize the need for a handover. In a separate paper, we discuss that potential of using the DCS as a handover instrument.

A new paradigm for discharge summaries may be needed. This new paradigm for discharge summaries would include not only data, but would focus on clear communication. We envision the EHR being able to generate a dynamic, prepopulated discharge summary that includes data already stored within the EHR. The discharge summary would then be completed by clinicians adding brief narratives that clarify the medical decision making of the case and that provide guidance and a useful handover to the next provider. We are investigating automatic generation of portions of the DCS and teaching the DCS as a handover in our institution. Further research may be necessary to clarify the most valuable elements of medical reasoning, how to record patient preferences, and how to most efficiently gather and present this knowledge.

Acknowledgements

The authors gratefully acknowledge the statistical expertise and assistance of Xiangyang Ye, MS who helped with the adapted Charlson Comorbidity Index calculations. The authors gratefully acknowledge the assistance of Beth Sakaguchi, MA who helped with proofreading to improve the readability.

References


Table 1. Demographics and description of discharge summaries*

<table>
<thead>
<tr>
<th>Group</th>
<th>Fall - Teaching General Medicine</th>
<th>Spring – Teaching Service</th>
<th>Spring - Attending Service</th>
<th>Groups Aggregated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discharge Dates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female n, (%)</td>
<td>N = 122</td>
<td>N = 33</td>
<td>N = 33</td>
<td>95 (51%)</td>
</tr>
<tr>
<td>N</td>
<td>63 (52%)</td>
<td>15 (45%)</td>
<td>17 (52%)</td>
<td></td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td>59 ± 19</td>
<td>58 ± 18</td>
<td>52 ± 15</td>
<td>58 ± 18</td>
</tr>
<tr>
<td><strong>Length of Stay (days)</strong></td>
<td>5.4 ± 1.9</td>
<td>5.2 ± 3.3</td>
<td>4.5 ± 4</td>
<td>5.2 ± 6.2</td>
</tr>
<tr>
<td><strong>Charlson Comorbidity Index Score</strong></td>
<td>3.1 ± 2.4</td>
<td>3.8 ± 2.6</td>
<td>2.8 ± 2.6</td>
<td>3.2 ± 2.5</td>
</tr>
<tr>
<td><strong>PCP Identified</strong></td>
<td>41 (34%)</td>
<td>12 (36%)</td>
<td>11 (33%)</td>
<td>64 (34%)</td>
</tr>
<tr>
<td><strong>Discharged to Home</strong></td>
<td>57 (47%)</td>
<td>15 (45%)</td>
<td>11 (33%)</td>
<td>83 (44%)</td>
</tr>
<tr>
<td><strong>Discharge Summary Statistics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Length (words)</strong></td>
<td>735 ± 300</td>
<td>818 ± 342</td>
<td>808 ± 442</td>
<td>762 ± 336</td>
</tr>
<tr>
<td><strong>Days to Dictation</strong></td>
<td>1.9 ± 3</td>
<td>1.5 ± 2.9</td>
<td>3 ± 6.6</td>
<td>2 ± 3.9</td>
</tr>
</tbody>
</table>

*There were no statistically significant differences between the three groups using the chi-squared test for dichotomous variables and an ANOVA for continuous variables.
Table 2. Average number of medications and occurrences of medical reasoning (about medications) per discharge summary.

<table>
<thead>
<tr>
<th>Medication Status</th>
<th># of Medications</th>
<th># of Occurrences of Medical Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continued</td>
<td>7.9</td>
<td>2.9</td>
</tr>
<tr>
<td>Changed</td>
<td>1.2</td>
<td>0.8</td>
</tr>
<tr>
<td>New</td>
<td>2.3</td>
<td>1.8</td>
</tr>
<tr>
<td>Discontinued</td>
<td>2.1</td>
<td>1.6</td>
</tr>
<tr>
<td>Unknown</td>
<td>7.3</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Table 3. Correlation of reporting pending results between DCS and EDW (2x2 table for source of reported pending results).

<table>
<thead>
<tr>
<th></th>
<th>DCS</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>1 (0.5%) Match</td>
<td>106 (56%) Discrepancy</td>
</tr>
<tr>
<td>Absent</td>
<td>6 (3%) Discrepancy</td>
<td>1 (0.5%) Explicit Match</td>
</tr>
<tr>
<td></td>
<td></td>
<td>61 (32%) Implicit Match</td>
</tr>
</tbody>
</table>

*In 13 (7%) of Discharge Summaries, there were mixed discrepancies so that some pending results were missing from the EDW while other pending results were missing from the DCS. The DCS with an explicit match stated, “there are no pending results,” which matched the EDW query. The DCSs with an implicit match did not mention the presence or absence of pending results, though none were found in the EDW.

Figure 1. Failure modes for providing continuity of care in discharge summaries.
Determinants of Consumer eHealth Information Seeking Behavior

Ryan H. Sandefer, MA, CPHIT1, Bonnie L. Westra, PhD, RN, FAAN, FACMI1, Saif S. Khairat, PhD2, David S. Pieczkiewicz, PhD3, Stuart M Speedie, PhD, FACMI4
1University of Minnesota, Minneapolis, MN; 2 University of North Carolina, Chapel Hill, NC

Abstract

Patients are increasingly using the Internet and other technologies to engage in their own healthcare, but little research has focused on the determinants of consumer eHealth behaviors related to Internet use. This study uses data from 115,089 respondents to four years of the National Health Interview Series to identify the associations between one consumer eHealth behavior (information seeking) and demographics, health measures, and Personal Health Information Management (PHIM) (messaging, scheduling, refills, and chat). Individuals who use PHIM are 7.5 times more likely to search the internet for health related information. Just as health has social determinants, the results of this study indicate there are potential social determinants of consumer eHealth behaviors including personal demographics, health status, and healthcare access.

Introduction

Engaging consumers in managing their health as a key to solving the core problems of healthcare access, quality, safety, efficiency, and effectiveness is an assertion boldly outlined by a plethora of landmark publications, including Institute of Medicine (IOM) reports.1,6 The import of these problems is exacerbated by out-of-control spending in health care that is expected to reach $4.6 trillion or 19.8% of the nation’s gross domestic product by 2020.7 Moreover, the already strained health care system will face an influx of patients as 30 million additional Americans obtain health insurance.8 Chronic health conditions are the leading cause of death and consume 75% of health care dollars in the U.S.9 Individuals with chronic conditions are more likely to be older, experience medical emergencies, and have higher healthcare expenses than other adults.10 The socioeconomic differences in health outcomes have been defined as the "social determinants of health."11

Multiple federal health policies and financial programs require or substantiate the need for consumers to engage electronically with their health care providers through the use of health information technologies (IT). The Centers for Medicare and Medicaid Services’ (CMS) Incentive Program for Meaningful Use of Electronic Health Record (EHR) requires health care providers to give consumers access to their health information and in future stages must allow submission of patient-generated information into EHRs.12,13 Emerging programs focused on care coordination and cost reduction also emphasize the use of patient portals and/or personal health records (PHRs), including the Accountable Care Organization (ACO) and the Patient Centered Medical Home (PCMH) programs.14

Recent research indicates that consumers and their families who use personal health information management (PHIM) applications such as PHRs have increased control over their personal health information and their use is associated with better health management, improved patient satisfaction, improved health care quality and safety, and more effective communication and collaboration between patients, doctors, caregivers and other stakeholders.15-19 Approximately 86% of US adults rate electronic access to PHRs and online health information as important.20 The use of PHRs is beneficial for patients and their providers in managing chronic diseases such as diabetes, hypertension, congestive heart failure, and depression.21-25 PHRs are also known to be beneficial for medication management,26-28 prevention screening,29, 30 clinical trial recruitment,31 patient education,32 and information exchange.33 There is a significant association between use of secure messaging with providers and lower Hemoglobin A1c values (i.e., glycemic control).34, 35 A greater frequency of secure email communication was associated with significant improvements in outcomes related to HbA1c screening and control, retinopathy screening, nephropathy screening, and blood pressure control; providers had an increased likelihood of improved performance on four diabetes measures.36

Patients are increasingly using the Internet as a source for obtaining health-related information. According to a study conducted by the Pew Research Center, 62.6% of Americans looked for health information online within the past year.20 An estimated 4.5% of all Internet searches are health related.37 Research has shown that there is variation in the quality of the information obtained from the Internet38 and the impact of the information obtained from Internet
searches on daily life is low. While a large number of Americans are using the Internet to search for health-related information, significantly fewer are adopting technologies such as PHRs to collect and store their personal health information and engage with healthcare providers.

There is, however, a mismatch between the perceived importance of access to PHIM tools versus their actual use. According to one estimate, only 7% to 9% of American adults used the Internet for tracking information in their personal health records. Use varies based on a variety of factors. PHR use is predominantly associated with people between the ages of 45 to 70. Racial disparity has been associated with patient portal use with higher use among Caucasians. Inconsistencies in Hispanic use of PHIM exist with the implication that further studies of culturally-diverse populations are needed. Females are reported to use PHRs more frequently than males. In a systematic review of PHR studies, Archer identified higher PHR use rates among individuals with chronic health conditions, those with frequent visits to health care providers, and families caring for older adults. Low income and low education levels are two additional factors that influence lower use of patient portals or other health IT for PHIM. Because PHIM is known to play an important role for those with chronic conditions, it is fundamentally important to identify the actual need and perceived value of PHIM among diverse users.

The use of the Internet for conducting health-related searches has been examined in terms of the socioeconomic factors, and research has shown that using the Internet for health-related searches is a more common among Americans than other PHIM activities. This is likely due to relatively few barriers, such as interdependencies with organizations, username and password issues, privacy and security concerns, and technology usability limitations. There has been limited research regarding the use of PHIM tools such as a PHR and its relationship to eHealth behaviors such as searching the Internet for health related information. The purpose of this study is to examine the association between those health-related Internet searches and use of at least one PHIM tool, including emailing healthcare providers, scheduling appointments online, refilling prescriptions using the Internet, and participating in health-related chat groups as well as other sociodemographic and health-related concerns.

**Methods**

This study used U.S. National Health Interview Series (NHIS) data that was collected by the Integrated Health Interview Series (IHIS). The Centers for Disease Control and Prevention conducts the NHIS annually “to secure accurate and current statistical information on the amount, distribution, and effects of illness and disability in the United States and the services rendered for or because of such conditions.” The NHIS randomly samples approximately 35,000 households and 87,500 persons annually, and the average response rate for the survey is near 90% of eligible households in the sample. Samples are drawn from each U.S. State and the District of Columbia and it is representative of the U.S. population. The NHIS has been conducted annually since 1957 and details about the NHIS have been published elsewhere. Data were extracted from the IHIS as a SAS text file, and the SAS text file was read into R and converted to a data set for all analyses.

The NHIS includes questions on a variety of health outcomes, demographic variables, and individual characteristics. The measure of information seeking as an eHealth behavior was the questionnaire item “During the past 12 months, have you ever used computers to look up health information on the Internet.” Responses were categorized into “Yes,” “No,” “Refused,” “Not Ascertained” or “Don’t Know.” For the purpose of the analysis, “No” was coded as 0 and “Yes” was coded as 1. The following variables were also included in the analyses: sex, gender, race, ethnicity, age, education level, health status, income, geographic region, care access, and hypertension prevalence. Hypertension was included in the analysis as a measure of a chronic condition for two reasons: there has been limited research examining PHIM among patients diagnosed with hypertension and hypertension is the most frequently reported chronic condition on the NHIS (47% of respondents reported hypertension). Education level was recoded into five categories (less than high school, high school or GED, some college or two year degree, bachelors degree, and graduate degree). Income was recoded into five categories (<35k$, $35-49.9k, $50-74.9k, $75-99.9k, and $100k+). The variable PHIM was created by combining together responses to questionnaire items related to the use of online health-related chat groups, email with healthcare providers, scheduling healthcare appointments online, and refilling prescriptions online. Responses were categorized into “Yes,” “No,” “Refused,” “Not Ascertained” or “Don’t Know,” and a response of “Yes” to any of these four questions resulted in a positive PHIM response. For the purpose of the analysis, “No” was coded as 0 and “Yes” was coded as 1. Any response of “Refused,” “Don’t Know,” or “Not Ascertained” were treated as missing values in the analysis.
For the purposes of this study, survey data regarding adults 18 years old or greater from 2009, 2011, 2012, and 2013 were combined. The 2010 survey was excluded since it did not address use of the Internet for seeking health information. Between 2009-2013, 492,948 individuals were interviewed across 191,395 households. The data used in this study was limited to adults responding “Yes” or “No” to the question regarding eHealth behavior, and was further limited to only individuals with complete data for the variables included in the logistic regression, resulting in a study sample size of 115,089 individuals. The University of Minnesota IRB deemed analysis of the NHIS data exempt from review.

In order to understand the characteristics of adults who engage in the eHealth behavior of information seeking, proportions and a Multivariable logistic regression were computed using the R version 3.1.0. Proportions were used to calculate the use of the Internet to search for health information across multiple variables. Multivariable logistic regression models were estimated to determine which variables were related to this eHealth behavior. Demographics, health and PHIM measures were included as independent variables in the logistic model. While PHIM activities are reported less frequently than using the Internet for health information seeking, the PHIM factor is included in this analysis to test the association between eHealth behavior and PHIM. The answer to the question about the use of the Internet to search for health information was the dependent variable. The results of the logistic regression are presented as odds ratios (OR) with the corresponding standard errors, 95% confidence intervals, coefficients, and Wald Statistic.

Results

Descriptive statistics are presented in Table 1. Overall, 45% of respondents reported using the Internet to search for health-related information but this ranged from 49.8% in 2009 to a low of 41.5% in 2012. Females were much more likely to search for health information than males (49.7% versus 39.8%, respectively). Adults were also more likely to search for health information if they were younger, white, highly educated, employed, and had higher incomes. Over 72.3% of adults reporting a graduate level education used the Internet to search for health-related information; 13.3% of adults with less than a high school degree reported conducting health-related searches. There were slight differences in geographic variation in health-related searching—the West and the Midwest had the highest rates (47.5% and 47.3%, respectively), while the Northeast and the South had the lowest (45.2% and 42.2%, respectively). Those individuals who have been diagnosed with hypertension reported searching for health information at a lower rate than those without hypertension (37.4% and 48.8%, respectively), as are those individuals who reported being in poor health compared to excellent health (24.9% and 51.7%, respectively). Adults who reported not having a usual place for medical care also reported searching for health information at lower levels (37.9%, and 46.6%, respectively).

Individuals who reported using the Internet for health-related searches also reported higher rates of PHIM activity. The overall rate of PHIM use is low—14.2% of respondents reported using at least one mode of PHIM such as use of online chats (3.6%), email with provider (5.9%), scheduling healthcare appointments online (4.7%), and refilling prescriptions online (6.7%). Individuals who reported using PHIM reported higher levels of using the Internet to search for health-related information than those who did not use PHIM (87.1% versus 37.9%).

Table 2 summarizes the output from the logistic regression on the eHealth behavior of using the Internet for health related searches. All the following results for each independent variable take into account adjustments for all of the other independent variables. Adults using PHIM were significantly more likely to use the Internet to search for health information (OR 7.50). Having a usual place for medical care was positively associated with using the Internet for health-related searches (OR 1.20). Adults reporting a diagnosis of hypertension are significantly more likely to conduct health-related searches (OR 1.08), and individuals who report being in Fair or Good Health were significantly more likely to conduct health-related searches (OR 1.22) than those reported being in poor health.

Adults who were younger, non-Hispanic, white, female, and had higher incomes were significantly more likely to conduct health-related searches. Adults who reported being in poverty were less likely to conduct health-related searches, but employment status was not significantly associated with conducting searches. These results are similar to previous research. Education level was positively associated with health-related searches. Adults with a graduate level education were over 9 times more likely to report a health-related searching than those with less than a high school degree.
## Table 1. Use of Internet for Looking Up Health Information by Respondent Characteristics*

<table>
<thead>
<tr>
<th>Demographics</th>
<th>% Respondents n = 115,089</th>
<th>% Using Internet to Search for HI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>55.1</td>
<td>49.7</td>
</tr>
<tr>
<td>Male</td>
<td>44.9</td>
<td>39.8</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-30</td>
<td>21.3</td>
<td>53.4</td>
</tr>
<tr>
<td>31-40</td>
<td>18.2</td>
<td>54.1</td>
</tr>
<tr>
<td>41-50</td>
<td>17.5</td>
<td>49.1</td>
</tr>
<tr>
<td>51-60</td>
<td>17.5</td>
<td>46.4</td>
</tr>
<tr>
<td>61-70</td>
<td>13.6</td>
<td>38.4</td>
</tr>
<tr>
<td>71+</td>
<td>12.1</td>
<td>17.9</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than High School</td>
<td>16.0</td>
<td>13.3</td>
</tr>
<tr>
<td>High School</td>
<td>25.6</td>
<td>31.1</td>
</tr>
<tr>
<td>Some College</td>
<td>30.7</td>
<td>52.5</td>
</tr>
<tr>
<td>Bachelors</td>
<td>17.8</td>
<td>66.5</td>
</tr>
<tr>
<td>Graduate</td>
<td>9.9</td>
<td>72.3</td>
</tr>
<tr>
<td><strong>Year</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>21.3</td>
<td>49.8</td>
</tr>
<tr>
<td>2011</td>
<td>25.5</td>
<td>44.6</td>
</tr>
<tr>
<td>2012</td>
<td>26.5</td>
<td>41.5</td>
</tr>
<tr>
<td>2013</td>
<td>26.6</td>
<td>45.9</td>
</tr>
<tr>
<td><strong>Income</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$35k$</td>
<td>41.5</td>
<td>31.7</td>
</tr>
<tr>
<td>$35-49.9k$</td>
<td>14.7</td>
<td>43.4</td>
</tr>
<tr>
<td>$50-74.9k$</td>
<td>17.0</td>
<td>51.7</td>
</tr>
<tr>
<td>$75-99.9k$</td>
<td>10.4</td>
<td>58.0</td>
</tr>
<tr>
<td>$100k +</td>
<td>16.5</td>
<td>66.0</td>
</tr>
<tr>
<td><strong>Poverty</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Poverty</td>
<td>18.4</td>
<td>28.9</td>
</tr>
<tr>
<td>Not in Poverty</td>
<td>81.6</td>
<td>48.9</td>
</tr>
<tr>
<td><strong>Employed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>40.7</td>
<td>34.2</td>
</tr>
<tr>
<td>Yes</td>
<td>59.3</td>
<td>52.8</td>
</tr>
<tr>
<td><strong>Geography</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Midwest</td>
<td>21.4</td>
<td>47.3</td>
</tr>
<tr>
<td>Northeast</td>
<td>16.2</td>
<td>45.8</td>
</tr>
<tr>
<td>South</td>
<td>36.7</td>
<td>42.2</td>
</tr>
<tr>
<td>West</td>
<td>25.7</td>
<td>47.5</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>75.3</td>
<td>47.4</td>
</tr>
<tr>
<td>Black/AA</td>
<td>15.6</td>
<td>34.6</td>
</tr>
<tr>
<td>AI/AN</td>
<td>1.0</td>
<td>32.8</td>
</tr>
<tr>
<td>Asian</td>
<td>6.2</td>
<td>45.7</td>
</tr>
<tr>
<td>Multi-Race</td>
<td>1.8</td>
<td>50.9</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>82.3</td>
<td>48.4</td>
</tr>
<tr>
<td>Hispanic</td>
<td>17.7</td>
<td>30.4</td>
</tr>
<tr>
<td><strong>Health Measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have a Usual Place of Care</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>83.9</td>
<td>46.6</td>
</tr>
<tr>
<td>No</td>
<td>16.1</td>
<td>37.9</td>
</tr>
<tr>
<td>Hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>68.3</td>
<td>48.8</td>
</tr>
<tr>
<td>Yes</td>
<td>31.7</td>
<td>37.4</td>
</tr>
<tr>
<td>Health Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>3.6</td>
<td>24.9</td>
</tr>
<tr>
<td>Fair</td>
<td>11.4</td>
<td>31.0</td>
</tr>
<tr>
<td>Good</td>
<td>58.7</td>
<td>46.3</td>
</tr>
<tr>
<td>Excellent</td>
<td>26.4</td>
<td>51.7</td>
</tr>
<tr>
<td>eHealth Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer PHIM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>85.2</td>
<td>37.9</td>
</tr>
<tr>
<td>Yes</td>
<td>14.2</td>
<td>87.1</td>
</tr>
</tbody>
</table>

*Except Geography (West & Midwest) and Age (18-30 & 31-40) all proportions are significantly different at p<0.01.
Table 2. Use of Internet for Looking Up Health Information by Respondent Characteristics

<table>
<thead>
<tr>
<th>Demographics</th>
<th>β</th>
<th>SE β</th>
<th>Wald Stat.</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td>0.57*** (0.56-0.59)</td>
</tr>
<tr>
<td>Male</td>
<td>-0.560</td>
<td>0.015</td>
<td>-38.402</td>
<td></td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-30</td>
<td>Ref.</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>31-40</td>
<td>-0.201</td>
<td>0.022</td>
<td>-9.015</td>
<td>0.82*** (0.78-0.85)</td>
</tr>
<tr>
<td>41-50</td>
<td>-0.493</td>
<td>0.023</td>
<td>-21.311</td>
<td>0.61*** (0.58-0.64)</td>
</tr>
<tr>
<td>51-60</td>
<td>-0.679</td>
<td>0.024</td>
<td>-28.259</td>
<td>0.51*** (0.48-0.53)</td>
</tr>
<tr>
<td>61-70</td>
<td>-1.089</td>
<td>0.027</td>
<td>-39.841</td>
<td>0.34*** (0.32-0.36)</td>
</tr>
<tr>
<td>71+</td>
<td>-1.983</td>
<td>0.033</td>
<td>-59.757</td>
<td>0.14*** (0.13-0.15)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than High School</td>
<td>Ref.</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td>0.750</td>
<td>0.027</td>
<td>27.365</td>
<td>2.12*** (2.01-2.23)</td>
</tr>
<tr>
<td>Some College</td>
<td>1.422</td>
<td>0.027</td>
<td>53.862</td>
<td>4.23*** (4.01-4.46)</td>
</tr>
<tr>
<td>Bachelors</td>
<td>1.883</td>
<td>0.030</td>
<td>62.449</td>
<td>6.57*** (6.20-6.97)</td>
</tr>
<tr>
<td>Graduate</td>
<td>2.197</td>
<td>0.035</td>
<td>62.466</td>
<td>8.99*** (8.40-9.64)</td>
</tr>
<tr>
<td><strong>Year</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>-0.289</td>
<td>0.021</td>
<td>-13.974</td>
<td>0.75*** (0.72-0.78)</td>
</tr>
<tr>
<td>2012</td>
<td>-0.455</td>
<td>0.021</td>
<td>-22.145</td>
<td>0.63*** (0.61-0.66)</td>
</tr>
<tr>
<td>2013</td>
<td>-0.247</td>
<td>0.021</td>
<td>-12.037</td>
<td>0.78*** (0.75-0.81)</td>
</tr>
<tr>
<td><strong>Income</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$35k$</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>$35-49.9k$</td>
<td>0.165</td>
<td>0.024</td>
<td>7.014</td>
<td>1.18*** (1.13-1.24)</td>
</tr>
<tr>
<td>$50-74.9k$</td>
<td>0.277</td>
<td>0.023</td>
<td>12.619</td>
<td>1.32*** (1.26-1.38)</td>
</tr>
<tr>
<td>$75-99.9k$</td>
<td>0.327</td>
<td>0.027</td>
<td>12.192</td>
<td>1.39*** (1.32-1.46)</td>
</tr>
<tr>
<td>$100k+</td>
<td>0.431</td>
<td>0.025</td>
<td>17.260</td>
<td>1.54*** (1.47-1.62)</td>
</tr>
<tr>
<td><strong>Poverty</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Poverty</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>In Poverty</td>
<td>-0.247</td>
<td>0.024</td>
<td>-10.419</td>
<td>0.78*** (0.75-0.82)</td>
</tr>
<tr>
<td><strong>Employed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>0.021</td>
<td>0.017</td>
<td>1.224</td>
<td>1.02 (0.99-1.06)</td>
</tr>
<tr>
<td><strong>Geography</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Midwest</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>-0.001</td>
<td>0.023</td>
<td>-0.054</td>
<td>1.00 (0.96-1.05)</td>
</tr>
<tr>
<td>South</td>
<td>-0.049</td>
<td>0.019</td>
<td>-2.578</td>
<td>0.95** (0.92-0.99)</td>
</tr>
<tr>
<td>West</td>
<td>0.100</td>
<td>0.021</td>
<td>4.728</td>
<td>1.11*** (1.06-1.15)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Black/AA</td>
<td>-0.573</td>
<td>0.021</td>
<td>-27.317</td>
<td>0.56*** (0.54-0.58)</td>
</tr>
<tr>
<td>AI/AN</td>
<td>-0.326</td>
<td>0.074</td>
<td>-4.407</td>
<td>0.72*** (0.62-0.83)</td>
</tr>
<tr>
<td>Asian</td>
<td>-0.696</td>
<td>0.030</td>
<td>-22.874</td>
<td>0.49*** (0.47-0.52)</td>
</tr>
<tr>
<td>Multi-Race</td>
<td>0.046</td>
<td>0.052</td>
<td>0.888</td>
<td>1.04 (0.94-1.15)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>0.606</td>
<td>0.021</td>
<td>28.804</td>
<td>1.83*** (1.76-1.91)</td>
</tr>
<tr>
<td><strong>Health Measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have a Usual Place of Care</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>No Usual Place</td>
<td>0.182</td>
<td>0.020</td>
<td>9.001</td>
<td>1.20*** (1.15-1.25)</td>
</tr>
<tr>
<td>Usual Place</td>
<td>0.079</td>
<td>0.017</td>
<td>4.46</td>
<td>1.08*** (1.05-1.12)</td>
</tr>
<tr>
<td><strong>Hypertension</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.151</td>
<td>0.048</td>
<td>3.139</td>
<td>1.16** (1.06-1.28)</td>
</tr>
<tr>
<td><strong>Health Status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Fair</td>
<td>0.201</td>
<td>0.045</td>
<td>4.488</td>
<td>1.22*** (1.12-1.34)</td>
</tr>
<tr>
<td>Good</td>
<td>0.054</td>
<td>0.047</td>
<td>1.158</td>
<td>1.07 (0.96-1.16)</td>
</tr>
<tr>
<td><strong>eHealth Measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer PHIM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>2.014</td>
<td>0.025</td>
<td>79.195</td>
<td>7.50*** (7.13-7.81)</td>
</tr>
</tbody>
</table>

**p<0.01, ***p<0.001**
Discussion

After examining the impact of social determinants of health, our findings show that individuals who report engaging in consumer eHealth applications by using online chat groups to learn about health topics, communicating with healthcare providers using email, scheduling an appointment with a healthcare provider using the Internet, and refilling prescriptions on the Internet are nearly 7.5 times more likely to search for health-related information online. This finding suggests that there is an association between PHIM and general engagement in healthcare information seeking. Given that consumer eHealth is negatively impacted by a variety of social determinants of health, the promotion and use of a single consumer eHealth modality (searching the Internet for health information) may promote wider adoption of consumer eHealth (PHIM). Considering the strong relationship between consumer eHealth engagement and the use of the Internet for conducting health related searches, healthcare providers should be encouraging their patients to use the Internet as this may be an entry point for patients into using PHIM.54 Individuals who email healthcare providers, schedule appointments online, and request prescription refills online are significantly more likely to search for health-related information online, and individuals who are more engaged with consumer eHealth report better health outcomes, patient satisfaction, and other measures of healthcare quality.19,30 This finding validates previous research demonstrating patient health status, experience, and knowledge of the Internet has a significant impact on seeking information online.55

Overall, our findings indicate that engaging in the eHealth behavior of using the Internet for health-related searches is much lower than has been previously reported (45% versus 63%),56 and that rate has been lower than 50% from 2009 to 2013. Previous estimates by the Pew Internet Research Project18 were based on a telephone survey with a much lower response rate than the NHIS survey 38 (18.1% and 75.5%, respectively), which potentially resulted in non-response bias. Our findings indicate that up through 2013 the level of adoption of these consumer eHealth practices has been generally low (14.2%).

Our research demonstrates a relationship between what have been labeled by the World Health Organization as the “social determinants of health” and using the Internet to conduct health-related searches.11 Individuals with a lower socio-economic status, lower levels of education, and non-White race or Hispanic ethnicity are less likely to engage in the eHealth behavior of using the Internet to search for health information, which confirms previous research.23,30,45,46,47

Our findings do indicate that individuals with hypertension (a chronic condition) are slightly more likely to conduct health-related Internet searches than individuals without hypertension, thus validating previous research.40,47 Individuals reporting a “Good” health status were also slightly more likely to conduct health-related searches. These findings suggest that those populations in poor health are somewhat less likely to engage in conducting health-related Internet searches. Indeed, this study provided additional evidence for a digital divide in consumer eHealth, implying that programs aimed at engaging patients through technology will struggle to reach underserved populations.

Our results also show that individuals who do not have a usual place for medical care are somewhat less likely to conduct a health-related Internet search. This demonstrates that access to medical care—a social determinant of health—is also a possible determinant of consumer eHealth behavior. The use of consumer eHealth applications, including patient-provider email, has been promoted as a tool for improving understanding and therefore decision making between patients and providers and increasing the level of patient and family engagement generally in healthcare.

Our results also suggest that patients who have a healthcare “home,” a place where they routinely obtain healthcare, are somewhat more likely to engage in searching the Internet for health information. This result supports recent initiatives aimed at promoting consumer eHealth, including the meaningful use requirement that 5% of unique patients have communicated with a healthcare provider through an online portal.13

There are several limitations with this study. First, the data was limited to self-reported responses regarding using the Internet to search for health information and use consumer eHealth applications. There are no additional details regarding the types of information searched or the frequency of searching by individual respondents, nor do we have details about frequency of emailing providers, scheduling appointments, refilling prescriptions, or participating in online chat groups.
Additional research is needed to evaluate the factors that drive adoption of consumer eHealth by individuals least likely to use the Internet for conducting health-related searches, and research is also needed to assess trends in consumer eHealth adoption by specific groups, including more detailed analysis of PHIM use by individuals who report chronic conditions. Research is also needed regarding the means by which individuals are conducting these searches and accessing eHealth technologies, including the use of mobile devices and mobile applications. This research could be useful for identifying best practices for engaging patients in personal health information management, and developing tools to assist healthcare organizations and individual healthcare professionals meet patient needs.

**Conclusion**

This study has demonstrated that there is significant relationship between the use of the Internet for conducting health-related searches and the use PHIM. This engagement is, however, impacted by a variety of demographic and health-related factors—“determinants of consumer eHealth.” Overall, the adoption of PHIM remains low, and to be successful, consumer health information technology must be designed to address the needs of people of varying ages, socio-economic status and geographic location. While many consumers are ready and willing to access their health information online, much less is known about consumer motivations related to personal health information management. The perspectives of a variety of consumers from diverse population groups are requisite to gaining an improved understanding of their personal health information management needs so that technologies that effectively capture and sustain their interest in managing their own health can be effectively designed and deployed.

**References**

3. Wunderlich G and Kohler P. Improving the Quality of Long-Term Care (Institute of Medicine Committee on Improving the Quality of Long Term Care). *Division of Health Care Services, Washington, DC*. 2001.


50. NHIS. National Health Interview Survey. CDC/National Center for Health Statistics, 2012.


Smartphone Data in Rheumatoid Arthritis – What Do Rheumatologists Want?

Phillip Say, MD, MS¹, Daniel M. Stein, MD, PhD¹, Jessica S. Ancker, MPH, PhD¹
Cheng-Kang Hsieh, MS², J.P. Pollak, PhD², Deborah Estrin, PhD¹,²
¹Weill Cornell Medical College, New York, NY, ²Cornell Tech, New York, NY

Abstract

Objective: To create a relevant and clinically informative visualization of passively collected patient mobility data from smartphones of rheumatoid arthritis (RA) patients for rheumatologists.

Methods: (1) Pilot analysis of smartphone mobility data in RA; (2) Assessment of rheumatologists’ needs for patient data through semi-structured interviews; and (3) Evaluation of the visual format of the RA data using scenario-based usability methods.

Results: We created a color-scale mobility index superimposed on a calendar to summarize the passive mobility measures from the smartphone that the rheumatologists confirmed would be clinically relevant.

Conclusion: This assessment of clinician data needs and preferences demonstrates the potential value of passively collected smartphone data to resolve an important data question in RA. Efforts such as these are necessary to ensure that any smartphone data that patients share with their doctors will not exacerbate clinician information overload, but actually facilitate clinical decisions.

I. Introduction

Informatics problem

Sixty-four percent of American adults owned a smartphone as of October 2014¹, which has significant implications in healthcare. Smartphones are intimately associated with so many daily activities that, if the data could be captured and analyzed, they could provide detailed information about an individual’s behavior as it relates to health. The smartphone itself passively tracks daily movements but it also allows for ease of data input into symptom or medication logs. There are accessories that can measure blood pressure, heart rate, oxygen saturation, expiratory flow velocity, cardiac electrical activity, and even brain electrical activity². A measure of how pervasive the belief is that these data sources could change healthcare delivery and research is the activity of several tech giants in this space, as exemplified by Google Inc.’s “Fit” and Apple Inc.’s HealthKit and recently announced ResearchKit technologies³.

Smartphone data gathered from patients with certain medical conditions, if processed and presented to their physicians effectively, could improve physician decision-making⁴. Unfortunately, physicians are already on the brink of information overload⁵. The average primary care physician might receive 1000 test results per week⁶ and is expected not only to be aware of the results, but also to analyze them and take the appropriate management actions. Data from patient smartphone applications, however potentially useful, is likely to be met by resistance from physicians if the clinical relevance is not immediately apparent. A significant amount of time and research must be spent on which measurements are truly useful and determining the best way to present the smartphone data in a clinically relevant and actionable manner.

Building on these ideas, our research group is studying methods for combining passive mobility monitoring and active questionnaires to create clinically informative measures for patients with rheumatoid arthritis (RA). As part of this larger effort, we sought to determine how to process and present smartphone data in a cognitively manageable, clinically relevant format for rheumatologists caring for patients with RA.
Rheumatoid arthritis management challenges

To place this work in context, it is helpful to recognize that RA management is fraught with many challenges related to assessing rapidly changing patient status and making appropriate decisions on the basis of this data. Managing an RA patient can be conceptualized as operating a complex feedback control system where the input variable is the RA medication and the output variable is the RA disease activity. Rheumatologists choose the dosing and frequency of non-steroidal anti-inflammatory drugs (NSAIDs) and disease-modifying antirheumatic drugs (DMARDs) such as methotrexate and steroids based on observing the RA disease activity. RA disease activity is characterized by painful, stiff joints in the hands and/or feet accompanied by fatigue and morning stiffness that takes at least an hour to resolve with activity. RA disease activity is unpredictable with flares that last a couple of days to several weeks. Clinical practice guidelines were designed based on available research to help rheumatologists improve their efforts at preventing long-term joint destruction and disability. The guidelines emphasize that rheumatologists frequently monitor patient disease activity and adjust medications and doses in order to achieve “tight control” of disease activity, which means maintaining remission or low disease activity at all times.

The guidelines encourage the use of composite disease activity indexes, which simplify and standardize the measurement of disease activity by assigning a single numerical score to patients for classification into one of four states: remission, low disease activity, moderate disease activity, or high disease activity. The composite disease activity indexes are well validated but have significant drawbacks. The composite disease activity indexes were created to capture the heterogeneity of RA into a standard measure, but it is possible for two patients with the exact same score to have vastly different signs and symptoms and actually be in two very different disease states. There is evidence suggesting a disconnect between the composite disease activity indexes and patient satisfaction with their treatment, which rheumatologists must overcome with their patients in order to make the best shared treatment decisions. It is also thought that most U.S. rheumatologists do not routinely use composite disease activity indexes in clinical practice because they are not practical to assess given the limited amount of time they have available during their patient encounter. These challenges of RA management call for more patient-driven data and a supplement to the composite disease activity index.

Objectives

The objective of this study is to create a relevant and useful visualization of patient generated smartphone mobility data and to elicit rheumatologist perspectives on 4 questions: (1) Is smartphone mobility data a relevant measure of RA disease activity? (2) Does this data answer clinician data needs in RA management? (3) Does it ease challenges of RA management? and (4) How would rheumatologists use it in the office setting?

2. Methods

This study had three steps: (1) Pilot analysis of smartphone mobility data in RA; (2) Patient data needs assessment through semi-structured interviews with rheumatologists; and (3) Evaluation of the visual format of the RA data using scenario-based usability methods.

Collecting sample mobility data from an RA patient

To obtain realistic data for the study, we recruited an RA patient to participate in data collection. The subject met the following inclusion criteria: 25 years old or older; worked full-time outside the home; and routinely carried a smartphone whenever leaving the house. The subject downloaded a commercially available motion/activity tracking software application onto an iOS smartphone, and was instructed how to link the application to the sensing technology platform used in this study. The subject provided information about which days were workdays versus non-workdays and notified the researcher by email when an RA flare occurred.

The smartphone recorded mobility measurements over time, leveraging the built-in accelerometer and GPS/Wi-Fi locator. We processed the raw data into multiple dependent variables for each day. They included: (1) time spent walking/running; (2) gait speed for walking/running; (3) distance travelled walking/running; (4) time spent using transportation and distance travelled; (5) maximum gait speed in miles per hour when walking for more than five minutes; (6) number of hours away from the house; (7) time of day that the subject left the house; and (8) geodiameter (furthest distance travelled from home). The rheumatologists reviewed these variables and gave
opinions about the usefulness of each of these and which ones might deserve a higher priority. The pilot data with the dependent variables measured by the smartphone were separated into workday and non-workday groups. Deciles for each dependent variable were calculated using Stata (version 13.1, StataCorp, College Station, TX). The decile values were used as a basis for creating a scoring scale from 0 to 9 unique for each variable.

Assessing clinical data needs of clinicians treating RA

Physicians with experience treating RA were recruited from a nationally known orthopedic hospital in New York, NY. Two of the physician subjects were co-investigators of a study related to this project. A snowball sampling approach was utilized in which these two co-investigators suggested the names of other rheumatologists for the interview.

A series of semi-structured interviews were conducted with these clinicians. The interview questions were developed based on literature review and designed with the following objectives: (1) to understand how rheumatologists assess an RA patient’s symptoms and functional disabilities; (2) to understand how they cognitively process this information to make medical management decisions; and (3) to identify gaps in data needs that might be addressed by patient-generated mobility data.

The interviews were audio-recorded and transcribed. Analysis was done in an iterative manner guided by a grounded theory approach. The insight and feedback gained from the interviews was used to alter the questions, patient scenarios, and format for summarizing the smartphone mobility data before the next subject was interviewed.

Evaluating the clinical utility of RA patient mobility data

During the physician interviews, a scenario-based design approach was used to evaluate the clinical utility of patient-generated mobility data. The approach involved presenting multiple RA patient scenarios to the physician subjects and asking them to think out loud while they decided how to manage the patients. Such “think-aloud protocols” or “verbal protocols” are useful in tracing problem-solving and decision-making steps. The scenarios were presented in two parts. After reading the first part the physicians were asked to describe how they would manage the patient. Then, they were given the second part of the scenario, which involved the same patient with some new or different information involving a presentation of the data. The physicians were asked to describe how they would manage the patient again, given this new information. The purpose of this method was to determine what effect, if any, the mobility index had on management. The presentation of the mobility index was revised after each interview, and the most current iteration was used in the scenarios. Physician subjects were also asked to critique the mobility index. This approach helped guide a discussion about the best way to present patient-generated mobility data so that it would integrate smoothly into the rheumatologist’s typical process for evaluating and managing RA patients.

Ethical approval

The Weill Cornell Medical College institutional review board approved the study. All physicians gave oral informed consent, and the patient subject granted written informed consent.

3. Results

Five board-certified rheumatologists participated who all reported <50% but >10% of their patients have a diagnosis of RA.

Physician clinical data needs when treating patients with RA

Five general themes about data needs when managing RA surfaced during the semi-structured interviews.

Theme 1: An accurate and detailed history is essential for helping rheumatologists recognize the patterns that help them properly identify RA flares

Flares present in various ways, even for the same patient. Some flares present suddenly while others have a more
gradual onset. Some cause obviously swollen joints but no joint tenderness while others cause no swollen joints but exquisite joint tenderness. “For some patients it is mostly pain, for some it is stiffness, and then there can be fatigue overriding the whole thing.” [Physician 1] Even laboratory tests can be unreliable. There are frequently patients with swollen, tender joints from RA with no elevation of the erythrocyte sedimentation rate or C-reactive protein (sensitive serologic markers for RA).

It is not always obvious that a patient is experiencing a flare. Back pain, a common symptom for all adults, is a good example. Rheumatologists must have enough historical information to rule out muscle strain, joint trauma, fibromyalgia, gout, herniated discs, neuropathic pain, and even depression as a cause of the back pain.

A detailed history can help rheumatologists evaluate a patient more accurately. It can help them detect the subtle patterns that define an RA flare so that they can more quickly and confidently prescribe the appropriate treatment.

**Theme 2: An accurate and detailed history is essential for fine-tuning dosing to maintain the delicate balance between RA symptoms and side effects from RA medications.**

Managing each individual RA patient’s treatment “is a delicate, fine balance because a small increase in DMARD dosage can improve symptoms, but that same small increase can also cause serious side effects.” [Physician 4] This effort is further challenged by the fact that rheumatologists are not always aware of which doses a patient is actually taking. Many patients are allowed to self-dose their prednisone and do not necessarily keep an accurate account of what they took or when. Also, medication adherence is a common challenge because patients are often more concerned about the side effects of DMARDs than they are of their RA. Once a reliable medication history is established, then subtle changes in symptoms and mobility can be associated with changes in dosing so that the optimum doses for the individual patient can be established.

**Theme 3: Ambivalence about validated disease activity indexes in RA management**

The composite disease activity indexes are well-validated measures of disease activity and are the basis for most of the RA management algorithms published in the clinical practice guidelines. However, two of the physicians had problems with the indexes. One physician felt that the indexes did not correlate with disease activity enough to be worth the time it takes to do them. “I’m not sure how helpful I find it. This may or may not be right but I like to treat the patient more than the number of things.” Another physician was very confident in the part of the index based on physical exam findings, but not in the single self-report question for the patient - “How much pain and stiffness have you had in the past week on a scale of one to ten?” This question appeared too subjective to be valuable because some people tolerate pain more than others, which makes the score inaccurate. The question also did not differentiate RA pain from something like degenerative joint disease pain, which is common among RA patients. The participant described a common scenario:

> “Sometimes the patient will look very comfortable but answers 8/10 where 10 is the worst. I’ll have to ask them if they mean a constant 8/10 or is that the worst score but most of the time they are a 3/10. I’ll end up having to ask them to try to pick whatever number pops into their head or a type of average, which is obviously extremely subjective. It also only asks about the last week. If I saw them three months ago, or even a month ago, then I really want to know about more than just the past week.”

The four physicians who used the indexes (whole or in part) felt that they generally correlated well with their subjective and physical assessments. However, these four physicians made the point that the composite disease activity index was only a part of their total assessment and did not replace any other part of their assessment. It was perceived as especially good at providing a way to consistently compare how a patient is doing relative to how they were doing at their previous appointment.

**Theme 4: Self-reporting of symptoms during the clinical encounter often gives an incomplete picture**

Rheumatologists rely heavily on a patient’s self-report to determine the true level of disease activity especially since there is no gold standard for measuring disease activity in RA and there is such wide variability in presentation. Unfortunately, patient self-report can lead to misrepresentations of the truth in many instances. “It is common to
have a patient complain about their swollen finger but say nothing about their knee that is swollen the size of a grapefruit.” The swollen finger may attract attention because it is new, whereas the knee has been symptomatic for months.

Self-report can also be inaccurate when patients understate the severity of their symptoms. One participant stated that it is common for “patients to say that they are doing much better, their joints feels good, and they are feeling good. Then I examine them and they have really swollen joints and are really tender.” Some patients accept more pain than they should. One explanation is the so-called “adaptation effect” in many chronic pain patients, in which pain appears normal to the patient because it has been present for so long. Only increases in pain are noted. Some patients who improve on a new treatment regimen will say, “Wow, I realize that I really wasn’t doing great now that I know how I am supposed to feel.” Other patients understate the severity of their symptoms simply because they are stoic and are not comfortable complaining. While other patients may be doing so much better on their new regimen that they understate their residual symptoms out of the fear of being asked to change to a new medicine.

Conversely, some patients overstate their symptoms. They could be more sensitive or vocal about lower levels of pain than the average patient, or the behavior could be drug seeking. Another possibility for a disproportionately high level of perceived pain is fibromyalgia. “Fibromyalgia is a condition that I believe acts as an amplifier. Pain begets pain in this situation and any pain they experience seems to cause them to complain more than other patients with similar physical findings.”

Theme 5: Self-report data generated between visits may help RA management, but presents new challenges

All the physicians agreed that more information about the time between office visits would be helpful in managing RA.

“The reality is that there is so much we don’t know about RA. We aren’t even sure exactly what causes flares. As a result, each flare is a learning experience that is lost. If somehow we could record what happened to the patient leading up to the flare and how they responded to the doses of prednisone or NSAIDS that they took, then it would provide valuable knowledge on how to prevent another flare and how to manage the next one.”

This type of knowledge could help rheumatologists achieve “tight control” more rapidly with the benefit of preventing the long-term damage that keeps occurring as long as their disease activity is elevated.

Two of the five interviewed physicians asked some of their patients to keep a symptom diary. The three physicians who did not stated that they would not have time to read a diary. “If I asked every patient to do a diary, 90% wouldn’t do it and then 10% would make it their life.”

One physician subject thought the journal information would be valuable for helping her manage her patients’ conditions, yet she was concerned that tracking symptoms at home would be tedious and boring and that patients would tire of it. She also expressed concern that forcing patients to document their symptoms might make them overly aware of them.

Creation of Passive Mobility Measures

The physicians all agreed that an index modeled after the composite disease activity index would be an effective way to represent a summary of the passive mobility measures and we decided to call it a “mobility index”. When asked about the individual mobility measures, the majority felt that time spent walking, time away from the house, and gait were the most useful in defining the level of RA disease activity. Four of the rheumatologists agreed that gait would be one of the more sensitive measures of increased disease activity, so we decided to apply more weight to gait in the mobility index weighted gait. Less weight was attributed to the time away from the house because two rheumatologists raised concerns that it was not as meaningful as the other measures.

The resulting formula for a summary estimate of the degree of mobility for the day was:
Another formula needed to be created to give a weekly summary of a patient’s mobility level. It could not be calculated as a simple average because rheumatologists want the weekly score to be defined by the worst days. Flares can often resolve in 48 hours, so its effect on a workweek mobility index could be washed out by a simple average. As a solution to this problem, the following formula was designed to enhance the signal of a low peak in the daily score where $d$ equals the number of days in the workweek.

$$
\text{Mobility index}_{\text{workweek}} = \frac{\sum_{n=1}^{d} \text{workday score}_n + 2(\text{lowest workday score})}{d + 2}
$$

For the purposes of the current study, we included only workday data and excluded weekend data because there was a large number of scores of “0” resulting in scores that were not interpretable.

**Clinically meaningful visualization of the Mobility Index**

The physician subjects were asked specific questions about how they would like this type of data presented to them. All of the physicians wanted to be able to review this smartphone data at the beginning of their office visit with the patient. They all wanted the smartphone data to be in the form of a simple single-paged report. The physicians wanted the smartphone data to help them understand symptom and side-effect trends over appropriate time intervals.

For a patient who was experiencing a flare, they expected to see daily changes. For a patient who was initiating a new medication, they would expect to see weekly changes. For a stable patient, they wanted to see an even broader view with a month-to-month summary.

In order to accomplish these goals, the daily and workweek mobility indexes were assigned a color shade according to its numeric value and the corresponding color shade was superimposed on a calendar view. Originally we used the colors red, orange, yellow, and green to represent different levels of mobility, but we determined that it was easier to convey the concept that these expressed ordinal categories of the same variable if we used different saturation levels of one color. That is, it was difficult to understand that green was a different level from red but easy to understand that dark purple is a different level from light purple. We used the color scheme in Figure 1. The darkest, most saturated level of purple was used to draw attention to the most impaired mobility. Figure 2 demonstrates the mobility index for the pilot patient with the color shades superimposed on the calendar for a daily score. Figure 3 demonstrates the weekly score.

![Figure 1](image.png)

**Figure 1.** Mobility index. The most impaired mobility is represented by the lowest mobility index values and the darkest shade of purple.
The physicians were generally enthusiastic about the potential benefits that mobility information on a calendar could bring to understanding RA after using it in the patient scenarios. “I think this type of information will be great. It will give me a better feel for the reality of what’s going on in their life,” one said. Another called it “a springboard for obtaining more insight into a person’s life than we get from the typical appointment interview.”

The physicians agreed that tracking a patient’s mobility would be most helpful in situations where the patients are understating their symptoms. In one of the scenarios where the smartphone data indicated that the patient was less mobile than they had admitted, one physician described how she could use this information as a way to point out that a patient may be accepting more disability than necessary and that there are changes in management that could help. “This information that is derived from objective measurements could also make it easier for patients to overcome their barriers of shame and embarrassment to reveal the truth about their symptoms,” another said.

The smartphone data did not have any effect on how the rheumatologists would treat a patient complaining of pain. The physicians agreed that they treat any patient who is in pain with maximal medical therapy. “Pain trumps everything,” one said. The physicians treat patients in pain based upon their self-report above all else, whether the smartphone data was in agreement or not with what the patient was saying. The mobility information also did not
affect treatment decisions for patients when it was consistent with the patient’s self-report and physical exam but it did offer an extra degree of confidence that they were doing the right thing. “I would use this mobility information as much as what was said and what I examined.”

Finally, the feedback from the clinician participants indicated that the mobility index in the calendar format could enhance the cataloging of RA signs and symptoms over time in the context of different environmental factors and medical interventions. “I ask patients to use a calendar to log symptoms and medications for me all of the time, but they don’t.” [Physician 3] Discussions with our clinician subjects suggested that successful adoption of an application like this by patients could help obtain generalizable knowledge about the behavior of RA. For example, if the mobility index and self-report are presented effectively for all of a rheumatologist’s RA patients, then after a while it could improve their understanding of the RA disease process enough to give them added knowledge to help them manage all of their patients. “If I have seen 10 people get 25% better in two weeks on a certain drug, then that is something I could share with a patient to give them an idea of what kind of benefits to expect from this new drug.”

4. Discussion

A key goal in creating mobile health-based tools for physicians is to assist them in making better patient management decisions. It is essential to have a thorough understanding of physicians’ data needs in order to create a tool that is relevant to clinical care. This also helps avoid the trap of developing a mobile health tool that is technologically sophisticated but does not have any real impact on the efficiency or quality of care delivered.

This study addressed those concerns by interviewing rheumatologists and describing the actual clinical problems and real world needs of rheumatologists. It also contributed information about the best format for presenting the smartphone data so that physicians would understand it and willingly adopt it into their patient management workflow.

The semi-structured interviews demonstrated that all of the rheumatologists were consistent in the opinion that more real-world data (generated outside of the clinical environment) about how RA affects an individual patient’s functionality would help them with management. Rheumatologists indicated that smartphone applications that track a patient’s movements throughout the day could be an indirect but valid measure of RA disease activity. Most importantly, all of the rheumatologists that were interviewed expressed a willingness to regularly spend a small amount of appointment time with their RA patients to review this type of data, as long as it was easy to understand, concise, and visual.

The patient scenarios demonstrated that visualizations of passive mobility were overall clinically useful to the rheumatologists. All of the five physicians understood the particular design that they were shown with little explanation. The calendar format added an improved understanding of the patient’s fluctuation of symptoms with time and helped physicians draw conclusions about trends. Allowing the calendar to change in granularity from daily to weekly to monthly helped the physicians adapt the data appropriately to the specific clinical situation. It also prompted rheumatologists to ask the theoretical patients more probing questions so that they would discover important details that the patient might have left out or forgotten. They felt that the addition of the visualizations would be better than performing a history and physical exam alone. It helped give them a more accurate picture of the patient’s true functionality, which even led to subtle changes in their choice of medical management.

There are numerous efforts ongoing to do this type of work in other domains. For Parkinson’s disease, researchers are using smartphones to transmit home motion-sensor data in a 3D animated form to clinicians for help with dose-finding16. For elderly patients, researchers are presenting physicians with a visualization of home motion sensor data and physiologic data to give them a better sense of patient overall health status17. There are also efforts to transform smartphone habit data with the Health Mashups System into a visual format that clinicians can evaluate and use to assist patients in promoting behavior change18. All of these projects aim to harness valuable information that the smartphone can collect and present it to physicians to aid in clinical decision-making.
Limitations

There are several limitations to this study. A single researcher performed data collection and analysis, although the iterative nature of the study meant that later physicians validated the interpretation of the analysis of the earlier interviews. The sample size was small, and the physician subjects were colleagues at a single medical center. The interviews revealed significant variability in preferences and disease management practices, but nevertheless, the results may not capture the diversity of opinion in other populations, such as rheumatologists in private practice or in different geographical areas. Given that many patients with RA will be cared for by a multidisciplinary team (physical and occupational therapists, nurses, etc.), another limitation of the study was that the clinicians interviewed were all physicians. Future work will include participation from other allied health professionals involved in the treatment of RA.

The pilot data collection was conducted with a single patient and may not reflect the full diversity of experiences of the RA population. An additional limitation is that we were unable to arrive at a stable score for non-workday readings and therefore excluded them from the current study. Additional data collection with more patients on non-workdays will help determine best how to calculate the meaningful score from non-workday data, as non-workdays may be even more informative about the severity of disease than workdays since most non-workday activities are voluntary. The current paper also does not address user testing, which is ongoing.

5. Conclusions

This assessment of clinician data needs and preferences demonstrates the potential value of passively collected smartphone data to resolve an important data question in RA, which is the daily activity level of the RA patient. This project also proposes a visualization solution identified by practicing rheumatologists as potentially valuable.

Concurrently with this data needs assessment, the mobile health development project is continuing. In addition to the passive data collection, we are developing a feature for patient self-report to allow patients to explain their schedule and report presence or absence of symptoms. Once all of this information can be captured simultaneously, then data analysis and data modeling will be able to improve and validate the choice of mobility measures, the formula for the mobility index, and the appropriate levels of severity to apply to the range of scores.

Patients are increasingly interested in capturing their health related data with their smartphones and wearable devices, and clinicians are going to be challenged in the future with interpreting these data. This study serves as an example of how to focus on a particular clinical problem, identify data needs, and design a data visualization technique that serves a clinical purpose. We think that this approach is necessary to ensure that the data patients generate and share with their doctors will not overwhelm them with information overload, but actually enhance their ability to provide the best possible care.

References


Vessel Delineation in Retinal Images using Leung-Malik filters and Two Levels Hierarchical Learning

Ehsan S Varnousfaderani, PhD¹, Siamak Yousefi, PhD¹, Christopher Bowd, PhD¹, Akram Belghith, PhD¹ and Michael H. Goldbaum, MD¹
¹Hamilton Glaucoma Center and the Department of Ophthalmology, University of California San Diego, La Jolla, CA, USA

Abstract
Blood vessel segmentation is important for the analysis of ocular fundus images for diseases affecting vessel caliber, occlusion, leakage, inflammation, and proliferation. We introduce a novel supervised method to evaluate performance of Leung-Malik filters in delineating vessels. First, feature vectors are extracted for every pixel with respect to the response of Leung-Malik filters on green channel retinal images in different orientations and scales. A two level hierarchical learning framework is proposed to segment vessels in retinal images with confounding disease abnormalities. In the first level, three expert classifiers are trained to delineate 1) vessels, 2) background, and 3) retinal pathologies including abnormal pathologies such as lesions and anatomical structures such as optic disc. In the second level, a new classifier is trained to detect vessels and non-vessel pixels based on results of the expert classifiers. Qualitative evaluation shows the effectiveness of the proposed expert classifiers in modeling retinal pathologies. Quantitative results on two standard datasets STARE (AUC = 0.971, Acc = 0.927) and DRIVE (AUC = 0.955, Acc = 0.903) are comparable with other state-of-the-art vessel segmentation methods.

Introduction
Color photography of the ocular fundus is a widely used imaging modality that permits the non-invasive analysis of the retinal microvasculature. Color images allow ophthalmologists and trained medical professionals to diagnose and monitor progression of retinal diseases, including age-related macular degeneration [1] and diabetic retinopathy [2]. Diabetic retinopathy as the primary cause of blindness can be prevented with treatment at an early stage, therefore WHO suggest yearly retina screening of patients. The reliable vessel segmentation can facilitate screening and improve vascular analysis that leads to improve Diabetic retinopathy detection [2, 3].

The manual analysis of ocular fundus images is time-consuming, expensive, and sometimes inaccurate. Computer aided ocular fundus image analysis permits the quantification of the extent of retinal abnormalities in vascular diseases and is fast and inexpensive and allows batch mode processing [3] [4]. Blood vessel segmentation is an important early step for the analysis of ocular fundus images. This can be a challenging task if the images are low quality and noisy vary in brightness, or have lesions underlying or adjacent to blood vessels [5]. Some automatic vessel segmentation methods extract feature vectors for pixels and use a set of manually segmented vessels (by experts) to train their classifiers, while the others are mainly developed based on filter response thresholding or other rule based techniques.

The matched filter technique [6] [7] is one of the earliest blood vessel segmentation methods that uses the maximum response of 12 different templates (2D kernel with Gaussian cross sections), in different orientations, to detect blood vessels. The filter parameters are then changed to increase the matched filter response to blood vessel detection by Al-Rawi [8]. The tracking approaches [9] [10] [11] build vessel trees by propagating vessel labels from manually or automatically selected pixels to unknown pixels, based on underlying correlations among pixels or tracking vessel center lines. The mathematical morphology is combined with cross-curvature evaluation and linear filtering by [12] [13] [14] to segment vessel-like patterns from background. The most recent filtering-based method is called B-COSFIRE [15], which uses a combination of shifted filter responses to extract vessels. B-COSFIRE is a trainable
filter approach in which the selectivity of a filter is determined automatically based on user prototype patterns such as straight vessels, bifurcations or a crossover points.

In learning based approaches, the proposed method by Ricci [16] uses a line detector on the inverted green channel of a retinal image. The average gray level on line passing through the pixels is computed in different orientations and the line with maximum value is selected. The line strength is computed as a difference between the selected line and the average of contrasts in the square neighborhood of the pixel. The line strength is high for vessel pixels and low for non-vessel pixels that allows unsupervised pixel classification by thresholding. Furthermore, an orthogonal line is defined based on a line of three pixels, centered on the midpoint of the main line and orthogonal to it. Then, its strength is obtained again by subtracting the average intensity on the orthogonal line from average intensity of the square neighboring of the pixel. The orthogonal line strength, line strength and intensity of pixels are combined to build a three-dimensional feature vector for every pixel. Later a SVM classifier is used to improve the supervised vessel classification result.

The proposed method by Fraz and et al [17] builds feature vectors based on orientation analysis of the gradient vector field, line strength measure, Gabor filter responses and morphological transformation for every pixel. Then, supervised methods, based on an ensemble of bagged and boosted decision trees, are used to classify pixels as vessels and non-vessels. The line elements are approximated based on grouped image ridges in the method developed by Staal [18]. The properties of line elements and their associate patch are used to build feature vectors for every pixel while the proposed method by Niemeijer [19] uses a response of a multiscale Gaussian filter to build feature vectors. Then, the k-Nearest Neighbor (kNN) classifier is used to classify pixels in both methods. Soares et al. [20] use a two-dimensional Gabor wavelet filter response, in multiple scales, to build feature vectors for every pixel. Then, a Bayesian classifier with class-conditional probability density functions (likelihoods) described as Gaussian mixtures is used to classify pixels as vessel and non-vessels. Most of the supervising method are designed to classify vessels and non-vessel regions regardless of the presence of abnormal pathologies such as lesions or exudates. Their performance is high when dealing with normal retinal images while it degrades drastically when dealing with abnormal retinal images.

The main goal of this paper is to develop a vessel segmentation method based on Leung-Malik filters [21] that seeks to minimize false positives and negatives in the face of retinal pathologies such as lesions, exudates and optic nerve abnormalities in diseased retinas. The proposed method uses a response of Leung-Malik filters in different scale and orientations to build set of feature vectors for every pixel. Then, a two level hierarchical learning framework is employed to segment vessels in abnormal retinal images. In the first levels, three classifier is trained to detect vessels, background and abnormalities then in the second level, a new classifier is trained to combine results of first level classification to delineate vessel and non-vessels. The performance of the proposed method on standard databases including STARE [7] and Drive [18] is comparable to the state of the art vessel segmentation methods.

The rest of the paper is organized as follows: in Section 2 we explain the proposed method and show how the LM filters responses can be used to detect blood vessels. In Section 3 we discuss and evaluate our method followed by discussions and finally we draw conclusions in Section 4.
The proposed method delineates vessels in 2D fundus images by classifying vessel or non-vessel pixels. The response of multi-scale, multi orientation Leung-Malik filter bank [21] is used to extract features for every pixel. The Leung-Malik filter bank consists of 60 filters, including first and second derivatives of Gaussians at 8 orientations and 3 scales (making a total of 48 filters); 8 Laplacian of Gaussian (LOG) filters; and 4 Gaussian filters. In our experiments the first and second derivative filters occur at the scale $\sigma = \{1, \sqrt{2}, 2\}$ with an elongation factor of 3 ($\sigma_x = \sigma, \sigma_y = 3\sigma_x$). The Gaussians occur at the four basic scales $\sigma = \{1, \sqrt{2}, 2, 2\sqrt{2}\}$ while the 8 LOG filters occur at scale $\sigma = \{1, \sqrt{2}, 2, 2\sqrt{2}, 3, 3\sqrt{2}, 6, 6\sqrt{2}\}$. The examples of the LM filter bank is shown in Fig1.b. The standard configuration for scales and number of orientations is used [21]. The maximum response of LM filters on the green channel image for different orientations, standard deviation of maximum responses on different scales and intensity value and its standard deviation in neighborhood of 5x5 of pixels are used to build a 14 dimensional feature vector for pixels as shown in Fig1.c. The LM filter bank extracts both textural, shape and intensity based features. The first and second derivatives of Gaussian in different orientation and scales have high response for elongated and vessel like objects and moreover the LOG and Gaussian filters have high responses for blob objects like exudate and hemorrhages. The extracted features in Fig1.c show how effectively the vessel like objects are highlighted in different orientation and scales using LM filters. Once feature vectors are extracted for every pixel then a Two Level Hierarchical Classification is applied to detect vessel and non-vessel pixels in abnormal retinal images as shown in Fig.2.
Fig 2. Two level Hierarchical classification framework. 

a) Expert classifiers to detect vessels, abnormalities and background pixels. 
b) Binary classification results and probability maps generated by expert classifiers. 
c) Local means and standard deviation of probability map as an input features for the classifier in the next level. 
d) Classifier in the level 2 and its outcome.
In the case of multi-class classification problem, the hierarchical classification framework takes advantages of simple expert classifiers in the first level to discriminate one class against all others. The expert classifiers are ensembles of n decision trees and they are trained to discriminate pixels in vessels, background and abnormal regions. Abnormalities in our experiments are considered as lesions, exudate, optic disk and regions with different textural pattern from normal retinal image. The expert classifiers are shown in Fig2.a.

The decision tree generates class label however the probability of samples originating from the class can be computed as the fraction of observations of the class in a tree leaf. The final probability map for expert classifiers is computed by averaging probability maps generated by all decision trees. The binary classification and probability maps of expert classifiers are shown in the Fig2.b.

The local mean and standard deviation in 3x3 square neighborhood in probability maps of three expert classifiers are used to build 6 dimensional feature vectors for every pixel. The 6 dimensional feature vectors are shown in Fig.2.c and it is used as an input for the classifier in the second level. In the second level, a new classifier based on ensembles of n decision trees is trained on 6 dimensional feature vectors (results of expert classifiers in the level one) to classify pixels as vessel or non-vessel. The classifier in the second level and its results are shown in Fig2.d. Different classifiers such as K nearest neighbor, Gaussian mixture models and ensembles of decision trees are used in the first and second levels but ensembles of 60 decision trees outperforms other classifiers. Therefore in our experiments the ensembles of 60 decision trees are used in the first and second level.

![Fig3. Illustration of Vessel segmentations. a) Green channel of retinal images from STARE and DRIVE datasets. b) Result of three expert classifiers in detecting background, abnormalities and vessels, c) segmented vessels by proposed method, d) Groundtruth vessel segmentations](image)

The performance of the proposed method is evaluated on standard STARE [7] and DRIVE [18] data sets. The STARE data set, collected at the Shiley Eye Center of UC San Diego, is used as a training set for the classifier. The STARE consists of 20 color retinal images (10 normal and 10 with pathology) captured by a Topcon TRV-50 fundus camera at 35 deg FOV. Two observers (an expert ophthalmologist, MHG) manually segmented images in which 10.4% and 14.9% of pixels are segmented as vessels by first and second observers respectively. In our experiment, segmentation by the first observer is used as gr ound truth as it is commonly used by other methods. In dealing with STARE data set a Cross-validation is used to train and test the classifiers on 75% and 25% images,
respectively. This process is repeated 4 times to test all images. The DRIVE data set consists of 40 images (20 for train and 20 for test) captured by a Canon CR5 nonmydriatic 3CCD camera at 45 field of view (FOV). The training and test sets are manually segmented one and two times respectively by observers trained by an ophthalmologist. In test set, 12.3% and 12.79% of pixels are segmented as vessels by first and second observers respectively and first observer is used as ground truth.

**Results and Discussions**

In qualitative evaluation, the performance of the proposed method in dealing with retinal images with abnormalities is evaluated. The green channel retinal images with and without abnormalities from STARE and DRIVE databases are shown in the Fig.3.a. The results of the three expert classifiers in detecting the image background, image abnormalities, and vessels are encoded with red, green and blue colors as shown in the Fig.3.b. The segmented vessels by the proposed method and the ground truth images are shown in the Fig.3.c, d.

The abnormalities close to the macula in the first STARE image are successfully detected by expert classifiers, but some part of their boundaries are wrongly considered as vessels, as highlighted with green and blue respectively in Fig.3.b. The expert classifiers successfully detect scattered and intensive pathologies in STARE images two and three, as shown in Fig.3.b. The proposed method takes advantage of expert classifiers and is able to segment vessels that are surrounded by confounding abnormalities, as shown for STARE images in the Fig.3.c. Fewer of the DRIVE images abnormalities, and result of vessel segmentation is comparable to ground truth labeling, (Fig.3.c, d). The regional post processing based on the shape and morphological characteristics of vessel or connecting disconnected vessels can improve the results, but in order to have fair comparison with other methods we do not use any post processing in our experiments.

The evaluation metrics of area under the curve (AUC) and accuracy (Acc) are used to quantitatively compare performance of the proposed method with state-of-the-art vessel segmentation methods on STARE and DRIVE databases, as shown in the Table.1. The accuracy of segmentation is computed as $Acc = \frac{TP+TN}{N}$, in which $TP$, $TN$ and $N$ are true positive, true negative and number of pixels respectively. To avoid having to select a single threshold for classification, one may scan through all possible thresholds, and observe the effect on the true positive rate and the false positive rate. The Area under the graphed curve is a reliable measure to compare performance of different methods. In our experiment the internal function (percurve) of Matlab software is used to estimate the AUC.

The proposed method achieves AUC of 0.971 on the STARE database, which is comparable with methods developed by Fraz [17], Marin [22] and Ricci [16] [22], with AUC of 0.977, 0.977 and 0.968 respectively. The performance of the proposed method is better on the STARE database than on the DRIVE database, because the STARE database has more retinal images with abnormalities.

<table>
<thead>
<tr>
<th>Table1. Performance of the proposed method and state-of-the-art methods on STARE and DRIVE databases with respect to AUC and ACC.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STARE</strong></td>
</tr>
<tr>
<td><strong>AUC</strong></td>
</tr>
<tr>
<td>Proposed Method</td>
</tr>
<tr>
<td>Staall et al. (2004) [18]</td>
</tr>
<tr>
<td>Soares et al. (2004) [20]</td>
</tr>
<tr>
<td>Al-Rawi et al (2007) [8]</td>
</tr>
<tr>
<td>Ricci and Perfetti (2007) [16]</td>
</tr>
<tr>
<td>Marin et al (2011) [22]</td>
</tr>
<tr>
<td>Fraz et al (2012) [17]</td>
</tr>
<tr>
<td>B-COSFIRE (2015) [5]</td>
</tr>
</tbody>
</table>

**Conclusion**

In this paper we have proposed a novel method to segment vessels in retinal images that have confounding abnormalities. The proposed method takes advantages of multiorientational Leung-Malik filters in different scales.
and makes use of a two-level hierarchical learning framework to detect vessels in diseased retinal images. The retinal abnormalities, vessels, and background are modeled by expert classifiers in the first level. The outcomes of the expert classifiers are combined to detect vessels in the second level. The qualitative evaluation shows that retinal abnormalities are successfully delineated from vessels and background by the expert classifiers. Moreover, the quantitative evaluation on two standard data sets, STARE (AUC = 0.971, Acc=0.927), DRIVE (AUC = 0.955, Acc =0.903), are comparable with state-of-the-art vessel segmentation methods.

References


Application of a Consumer Health Information Needs Taxonomy to Questions in Maternal-Fetal Care

Jared A. Shenson, B.S., Ebone Ingram, B.A., Nadja Colon, M.D., and Gretchen Purcell Jackson, M.D., Ph.D.
Vanderbilt University Medical Center, Nashville, Tennessee

Abstract

Pregnancy is a time when expectant mothers may have numerous questions about their unborn children, especially when congenital anomalies are diagnosed prenatally. We sought to characterize information needs of pregnant women seen in the Vanderbilt Children’s Hospital Fetal Center. Participants recorded questions from diagnosis through delivery. Questions were categorized by two researchers using a hierarchical taxonomy describing consumer health information needs. Consensus category assignments were made, and inter-rater reliability was measured with Cohen’s Kappa. Sixteen participants reported 398 questions in 39 subcategories, of which the most common topics were prognosis (53 questions; 13.3%) and indications for intervention (31 questions; 7.8%). Inter-rater reliability of assignments showed moderate (κ=0.57) to substantial (κ=0.75) agreement for subcategories and primary categories, respectively. Pregnant women with prenatal diagnoses have diverse unmet information needs; a taxonomy of consumer health information needs may improve the ability to meet such needs through content and system design.

Introduction

Pregnancy is a common health condition that generates a diverse set of consumer health information needs, not only for the mother, but also for all individuals who are involved in the care of the mother and child. In the United States, there are approximately 6.5 million pregnancies each year, resulting in over 4 million live births. Although the duration of pregnancy is limited, it is nonetheless a 40-week health condition with a variable and changing course that can be viewed within an illness trajectory framework. Pregnant women must adopt new health practices to support their developing baby, manage existing medical conditions in the context of pregnancy, and address maternal or fetal complications. For the numerous American women each year who are pregnant in addition to having one or more existing chronic illnesses, the concurrent management of these conditions can be quite challenging. Even women with relatively “normal” pregnancies may have widely varied experiences influenced by their individual symptoms, degree of physical limitations, and effects of their lives.

Annually in the United States, approximately 875,000 women experience one or more complications of pregnancy, and approximately 120,000 children are born with birth defects. When disruptions in the pregnancy trajectory, such as a maternal complication or fetal anomaly, are identified, the conceptualization and management of the pregnancy changes drastically, especially as more information is discovered and measures to affect maternal or fetal outcomes are initiated. Parents may be faced with difficult decisions about high-risk testing or treatments, pregnancy termination, or the withdrawal of support after the birth of a child with a poor prognosis. Maternal or fetal complications can result in newborns with chronic illnesses that will require lifelong management. During this time, the mother and her support network have increased interaction with the healthcare system and may experience significant physiological, psychosocial, and financial changes.

Obtaining information is the most basic and perhaps most important action one can take to manage a health condition and make informed decisions. Consumer health information needs have been studied across a number of patient populations, but little is known about the needs of pregnant women. Among those referred for genetic counseling, the most commonly expressed concerns were the risks of certain diseases and interpretation of genetic tests. Most questions in this population were prompted by uncertainty and the need for reassurance. Studies of prenatal consultations for congenital anomalies identified diverse information needs about the nature of the anomaly, treatment options, and prognosis. They also highlighted the value of these consultations in reducing anxiety, preparing the families, and maintaining hope. Parental and caregiver information needs and information-seeking behaviors are notably dynamic over the course of a disease and as a result of interventions performed, such as surgery.
Adequately addressing the information needs of parents and caregivers can improve prenatal care as well as maternal and fetal outcomes. It is well known that appropriate prenatal education and care reduces the risk of preterm birth, pregnancy complications, and congenital anomalies. Several long term studies of nurse home visits to educate pregnant women before and after delivery have shown dramatic improvements in a wide variety of social, economic, developmental, and other health-related outcomes for the mother, family, and child.

Patients and families are increasingly turning to the Internet and other health information technologies (e.g., patient portals and mobile applications) to manage health-related information and answer questions. Complex social, economic, and cognitive factors are likely to contribute to the willingness to use such technologies and individuals’ success in finding appropriate answers. Knowledge about information needs, information-seeking behaviors, and resource preferences are needed to guide the design and support the adoption of health information technologies with the goal of improving prenatal care and outcomes. In this study, we characterized the information needs of pregnant women seen in an advanced maternal-fetal care clinic using a taxonomy of consumer health information needs developed by our research team. We also evaluated the reliability of this taxonomy in describing consumer health information needs.

**Methods**

*Population and data collection*

We examined the quantity and nature of information needs of pregnant women evaluated for a pregnancy complication or congenital disorder at the Fetal Center of the Monroe Carell Jr. Children’s Hospital at Vanderbilt University Medical Center. Patients who were 18 years of age or older were approached to take part in this study, and those women providing informed consent were enrolled. Data were collected between November 2007 and January 2009. The research protocol was approved by the Vanderbilt University Institutional Review Board.

Participants were given notebooks and asked to record questions that arose regarding the medical problems of their unborn children from the time of diagnosis through the end of their pregnancies. Questions were collected from the notebooks at the time of follow up clinic visits and at the end of the study.

*Data analysis*

Participants’ journal questions were transcribed into text files with all identifiable information removed. Multi-part questions were separated into individual question segments for coding analysis. We categorized questions using a taxonomy we developed to represent consumer health information needs (Table 1). Model development started with a representation of Clinical Information needs, which are questions requiring clinical knowledge; this model of clinical information needs has been used to organize and facilitate search in online medical information resources. This model divides clinical knowledge into four primary categories: Problems, ranging from non-specific findings such as fever to well-defined diseases such as stage 4 melanoma; Management, the general framework for addressing a clinical problem; Tests, encompassing any diagnostic modality; and Interventions, including any therapeutic modality from education to medications or surgical procedures. Subcategories, such as indications or contraindications for tests and interventions capture more detailed types of information.

A research team comprised of clinicians, medical students, and human factors experts expanded the taxonomy after analyzing a variety of consumer health communications, including journal entries, patient-provider messages exchanged via a patient portal, and patient and family interviews about informational needs. After independent and consensus review of these sources, three additional primary categories were added to the taxonomy: Medical needs, addressing desire for or delivery of medical care; Logistical needs, pertaining to the pragmatic and supporting aspects of care delivery; and Social communication, expressions of social interaction or an interpersonal relationship that is not directly related to care delivery. Subcategories were then enumerated to describe the specific question types identified in these primary categories. Our final taxonomy contains seven primary categories, 59 subcategories, and one category (Other) for content that cannot be categorized due to being incomplete or incomprehensible.

Two raters with clinical experience were trained to use the taxonomy on the first 200 questions collected. The remaining questions were independently coded by the two raters, with their responses used to calculate measures of inter-rater reliability. Cohen’s kappa was used for this purpose as it estimates the beyond-chance agreement of two raters coding with mutually exclusive categories. A gold standard consensus for all messages in the corpus was developed through discussion between the two raters and the lead taxonomy author (GPJ).
<table>
<thead>
<tr>
<th>Category / Subcategory</th>
<th>Example Question (Participant #)</th>
<th>Count</th>
<th>Percent of Primary Cat.</th>
<th>Percent of All Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problems</td>
<td></td>
<td>146</td>
<td>–</td>
<td>36.7</td>
</tr>
<tr>
<td>Definition</td>
<td><em>What is gastroschisis? (52)</em></td>
<td>4</td>
<td>2.7</td>
<td>1.0</td>
</tr>
<tr>
<td>Epidemiology</td>
<td><em>How often are babies born with this defect? (52)</em></td>
<td>11</td>
<td>7.5</td>
<td>2.8</td>
</tr>
<tr>
<td>Risk Factors</td>
<td><em>Does holoprosencephaly increase the risk for sudden infant death syndrome? (20)</em></td>
<td>23</td>
<td>15.8</td>
<td>5.8</td>
</tr>
<tr>
<td>Etiology</td>
<td><em>What causes holoprosencephaly? (20)</em></td>
<td>12</td>
<td>8.2</td>
<td>3.0</td>
</tr>
<tr>
<td>Pathogenesis / Natural History</td>
<td><em>How does pregnancy with twins affect my blood sugar levels worse than a singleton? (31)</em></td>
<td>11</td>
<td>7.5</td>
<td>2.8</td>
</tr>
<tr>
<td>Clinical Presentation</td>
<td><em>How much intestine are usually outside? (45)</em></td>
<td>23</td>
<td>15.8</td>
<td>5.8</td>
</tr>
<tr>
<td>Differential Diagnosis</td>
<td><em>Why isn’t baby moving legs very much? (13)</em></td>
<td>4</td>
<td>2.7</td>
<td>1.0</td>
</tr>
<tr>
<td>Related Diagnoses</td>
<td><em>Are there other abnormalities that we should be looking for after this? (37)</em></td>
<td>5</td>
<td>3.4</td>
<td>1.3</td>
</tr>
<tr>
<td>Prognosis</td>
<td><em>What is the chance of survival after birth? (18)</em></td>
<td>53</td>
<td>36.3</td>
<td>13.3</td>
</tr>
<tr>
<td>Management</td>
<td></td>
<td>74</td>
<td>–</td>
<td>18.6</td>
</tr>
<tr>
<td>Goals / Strategy</td>
<td><em>What’s going to happen after baby is born? (52)</em></td>
<td>12</td>
<td>16.2</td>
<td>3.0</td>
</tr>
<tr>
<td>Tests</td>
<td><em>How common is genetic testing during pregnancy? (33)</em></td>
<td>5</td>
<td>6.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Interventions</td>
<td><em>How can this [gastroschisis] be fixed? (52)</em></td>
<td>21</td>
<td>28.4</td>
<td>5.3</td>
</tr>
<tr>
<td>Sequence / Timing</td>
<td><em>Will circumcision be done during his first surgery or after? (32)</em></td>
<td>21</td>
<td>28.4</td>
<td>5.3</td>
</tr>
<tr>
<td>Personnel / Setting</td>
<td><em>Can we deliver at the hospital back home or do we need to come to [Hospital]? (56)</em></td>
<td>15</td>
<td>20.3</td>
<td>3.8</td>
</tr>
<tr>
<td>Interventions</td>
<td></td>
<td>101</td>
<td>–</td>
<td>25.4</td>
</tr>
<tr>
<td>Definition</td>
<td><em>What is a shunt? (19)</em></td>
<td>3</td>
<td>3.0</td>
<td>0.8</td>
</tr>
<tr>
<td>Goals</td>
<td></td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Mechanism of Action</td>
<td><em>What’s the drug and what does it do? (35)</em></td>
<td>1</td>
<td>1.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Efficacy</td>
<td><em>Is it proven that there is a better outcome by having the surgery before the baby is born? (19)</em></td>
<td>3</td>
<td>3.0</td>
<td>0.8</td>
</tr>
<tr>
<td>Indications / Contraindications</td>
<td><em>Is this a reason to abort my child and not give him a chance (33)</em></td>
<td>31</td>
<td>30.7</td>
<td>7.8</td>
</tr>
<tr>
<td>Preparation</td>
<td><em>Do I need to be NPO before my C-section? (17)</em></td>
<td>2</td>
<td>2.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Technique / Administration</td>
<td><em>How long will his surgery take? (34)</em></td>
<td>16</td>
<td>15.8</td>
<td>4.0</td>
</tr>
<tr>
<td>Monitoring</td>
<td></td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Post-intervention Care</td>
<td><em>After his surgery, how long before he can have a bath? (32)</em></td>
<td>20</td>
<td>19.8</td>
<td>5.0</td>
</tr>
<tr>
<td>Advantages / Benefits</td>
<td><em>Will it be best to breastfeed or use formula? (23)</em></td>
<td>4</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Costs / Disadvantages</td>
<td></td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Adverse Effects</td>
<td><em>Will my baby hurt after surgery? (52)</em></td>
<td>21</td>
<td>20.8</td>
<td>5.3</td>
</tr>
<tr>
<td>Table 1 (continued)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td><strong>Tests</strong></td>
<td>23</td>
<td>–</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>Definition</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Goals</td>
<td>2</td>
<td>8.7</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Physiologic Basis</td>
<td>1</td>
<td>4.3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Efficacy</td>
<td>1</td>
<td>4.3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Indications /</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Contraindications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preparation</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Technique /</td>
<td>2</td>
<td>8.7</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Administration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation</td>
<td>14</td>
<td>60.9</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>Post-test Care</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Advantages /</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Benefits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Costs / Disadvantages</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Adverse Effects</td>
<td>3</td>
<td>13.0</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>Medical Needs</td>
<td>20</td>
<td>–</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>Scheduling</td>
<td>1</td>
<td>5.0</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Medical Equipment</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Personnel /</td>
<td>2</td>
<td>10.0</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Referrals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prescriptions</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Problem</td>
<td>2</td>
<td>10.0</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Follow Up</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td>4</td>
<td>20.0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Tests</td>
<td>1</td>
<td>5.0</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Interventions</td>
<td>10</td>
<td>50.0</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Logistical Needs</td>
<td>33</td>
<td>–</td>
<td>8.3</td>
<td></td>
</tr>
<tr>
<td>Contact Information</td>
<td>2</td>
<td>6.1</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Facility / Policies</td>
<td>25</td>
<td>75.8</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>Insurance /</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Billing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interventions</td>
<td>5</td>
<td>15.2</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>Medical Records</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Pers. Documentation</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Health IT</td>
<td>1</td>
<td>3.0</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Tests</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Social Communication</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>–</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

† All subcategories (Acknowledgement, Concerns, Relationship, and Miscellaneous) contained zero messages.
Results

Participants and journal questions

Sixty pregnant women were enrolled in this study. The age of the participants ranged from 18 to 38 years with a median age of 24.4 years. Fifty-five participants were Caucasian, two were black, and three declined to provide their race. The average gestational age at enrollment was 26 5/7 weeks with a median of 27 weeks and range of 14 to 37 5/7 weeks. The average time from enrollment to delivery was 67.1 days for all study subjects. There were 67 prenatal diagnoses in the fetuses of the 60 women who participated in the study. The most common anomalies were gastroschisis (n = 13), spina bifida (7), ventriculomegaly (7), and congenital diaphragmatic hernia (7). Other diagnoses included holoprosencephaly, Dandy Walker malformation, hydrocephalus, cleft lip or palate, hypoplastic left heart, hypoplastic right heart, transposition of the great vessels, congenital cystic adenomatoid malformation, pulmonary sequestration, multicystic or polycystic kidney, polyhydramnios, hydronephrosis, skeletal dysplasia, talipes, absent radius, anemia, thrombocytopenia, heterotaxy, trisomy, and conjoined twins.

Nineteen women returned notebooks for question analysis, and 16 of these participants submitted at least one question during the study period. The average age of patients returning notebooks was 24.6 years with a median age of 23 years. All patients returning notebooks were Caucasian. The average gestational age at enrollment for participants returning notebooks was 28 3/7 weeks with a median of 28 6/7 weeks. For the patients who submitted at least one question, the average time from enrollment to delivery was 58.3 (range of 1–130) days. Participants returning notebooks each reported up to three congenital anomalies associated with their pregnancies, including gastroschisis, ventriculomegaly, hydrocephalus, holoprosencephaly, microcephaly, heterotaxy, congenital adenomatoid malformation, congenital diaphragmatic hernia, trisomy 6, choroid plexus cyst, mosaicism, lemon sign (a fetal skull ultrasound finding), and hypoplastic left heart syndrome. The participants reported a total of 398 questions with an average of 21 questions per participant who returned a notebook.

Health information needs

The great majority (86.3%; Table 1) of questions submitted by participants concerned clinical information needs, with most related to Problems (36.7%), Interventions (25.4%), or Management (18.6%). Clinical information needs about Tests accounted for just 5.8% of all questions. The remainder of participants’ questions addressed Medical Needs (5.0%) or Logistical Needs (8.3%). One question could not be categorized as it did not contain a question or expressed need (Other, 0.3%). No questions were identified as Social Communication.

Participants’ questions most commonly addressed specific concerns about the prognosis of their fetus (13.3%), the indications or contraindications for performing particular interventions (7.8%), policies of the medical facility where the participant and their fetus was receiving care (6.3%), the clinical presentation of the problems affecting their fetus (5.8%), and the risk factors for the problem (5.8%). Sixty-six percent (39/59) of all subcategories in the taxonomy had at least one question meeting their criteria for coding.

Within primary categories, the most commonly identified subcategories were: fetal prognosis (36.3% of all Problem questions); recommended or available interventions and the sequence/timing of management plans (each 28.4% of all Management questions); indications or contraindications for particular interventions (30.7% of all Interventions questions); interpretation of test results (60.9% of all Test questions); requests for administration of interventions (50.0% of all Medical Needs questions); and facility policies (75.8% of all Logistical Needs questions).

Taxonomy reliability

Two raters independently assigned codes to a random sample of 198 questions that were distinct from those used to train the raters in the use of the taxonomy. The raters achieved 60.1% agreement with $\kappa = 0.57$ when using the subcategories only. This represents moderate agreement beyond chance, according to the widely used criteria proposed by Landis and Koch. When the raters’ subcategory codes were mapped to their respective primary categories, percent agreement improved to 82.3% with $\kappa = 0.75$, representing substantial agreement beyond chance.

Discussion

Health information needs of pregnant women

Through the application of a novel taxonomy of consumer health information needs, we have identified and categorized diverse needs expressed by women carrying high-risk pregnancies. The journaling process by which
participants generated and recorded their questions enabled participants to reflect and focus on the many clinical information concerns that arose as a result of their fetus’s prenatal diagnosis, and greater than 85% of questions addressed such clinical information needs. Many women wished to understand the implications of the diagnosis, including the prognosis of their fetus, the clinical features of the congenital anomaly or syndrome, and the natural history of the problem. There was also significant concern about the frequency of and reasons why the problem arose and the likelihood that it would affect future pregnancies (i.e., epidemiology, risk factors and etiology).

With many of the fetuses likely to require interventions including surgery and/or medications, women wanted to know more about available management strategies and timing as well as specific details of the interventions that might be required. Among questions oriented to a specific intervention, most dealt with indications, contraindications, adverse effects, care for the infant after the intervention, and the technique by which the intervention is administered. These findings reinforce the key informational elements that should be included in both verbal discussions between clinicians and patients as well as electronic resources (e.g., patient portals, knowledge websites) developed to support this patient population. Despite the importance of testing in evaluation and management of most prenatal diagnoses, few questions addressed such information needs and most focused on the interpretation of test results that the participants had already received. This observation suggests that clinicians and genetic counselors providing support to this population may need to give additional attention to explaining the results of testing performed and the implications for problem management.

Most non-clinical information needs were logistical in nature, focused on either facility policies (e.g., restrictions on visitors to the neonatal intensive care unit) or the availability of services or equipment at a particular medical facility. These unanswered questions are an easily addressed category of unmet information needs, and providing such information may help decrease anxiety in a stressful time for the expectant mother. Few active medical needs were expressed, but these frequently included the need for maternal interventions, such as scheduling a Cesarean section, treating swollen joints, or enrolling in special birthing classes.

As the questions in our study were collected throughout a participant’s pregnancy, we observed changing health information needs as the participants progressed toward delivery. This trend was demonstrated by Participant 32, who initially asked questions about future management of her fetus once an infant (e.g., “Should I make sure any kids he will be around aren’t sick?”), but later asked questions about the impending surgery for her fetus once born (e.g., “After his surgery how long before he can have a bath?”). Such changes reflect a woman’s evolving priorities and need for answers as more information is discovered and measures to affect maternal or fetal outcomes are planned or initiated. Our ongoing research projects are evaluating patterns of information needs over time and expanding the illness trajectory framework for pregnancy.

**Taxonomy development, reliability, and implications**

In addition to describing consumer health information needs among women with complicated pregnancies, our study reports the preliminary evaluation of a novel taxonomy for their organization. Notably, our evaluation demonstrates our taxonomy’s ability to comprehensively categorize a large number of questions with moderate to substantial reproducibility. Our study reports the distribution of self-reported information needs in pregnant women experiencing a problem with their pregnancies or unborn children. The types of questions identified in other patient populations and sources of consumer health information needs (e.g., patient portal messages) may differ significantly. While we did not identify questions for all categories in our proposed taxonomy in the current study, additional work by our group evaluating other sources of consumer information needs has shown a broad range of questions across the proposed categories and sufficient coverage to justify their inclusion. Additional validation on other sources of consumer health questions and an analysis of categories with high rates of disagreement between our raters will enable refinement of our taxonomy for future use.

Other researchers have proposed taxonomies for clinical questions asked by physicians, but these are complex and contain potentially overlapping categories. Consumers’ health information needs are more varied than physicians’ and extend beyond the clinical information domain. Classifying consumer health information has been a research priority, and some studies have examined patient and caregiver needs in specific diseases. However, comprehensive taxonomies to characterize broadly the semantic types of consumer information needs independent of domain or disease have not been proposed.

As clinicians, informaticists, and systems developers collaborate to build resources to support healthcare consumers, a comprehensive taxonomy of consumer health information needs may have several important applications. First, our taxonomy could be used to characterize the information needs of populations, as we have demonstrated here for
women carrying high-risk pregnancies. Such research may identify gaps in current sources of knowledge for consumers and guide content authors in producing resources that best address information needs. Our taxonomy could also be used to triage questions or map to appropriate information resources to provide answers. The broad classification of information needs into clinical information, medical, and logistical needs may guide selection of an appropriate resource type, and the specific subcategories in our taxonomy can provide filtering of content to meet the consumer’s particular need. For example, clinical information needs might be answered by medical textbooks or corresponding consumer health knowledge resources, while logistical needs could be directed to appropriate pages on the healthcare institution’s website. Medical needs might require urgent intervention by a healthcare provider, but social communications could be addressed during normal business hours.

Limitations

Our study was conducted in an advanced maternal-fetal care setting, and the information needs observed may not generalize to women with normal pregnancies. We did not collect information about whether the participants had been pregnant previously, and the median age of our participants of 23 years was slightly younger than the average age of first-time pregnancy in the United States of 26 years. Participants’ age and parity are likely to have affected the types of questions reported. We had a relatively low response rate, as dedicated research personnel were not available to collect notebooks at all clinic visits. The questions collected likely reflect the needs of patients with more complicated pregnancies, as healthy mothers may not have returned to the Fetal Center or our tertiary care center for delivery. Although only 16 participants provided questions, we obtained 398 questions for analysis, which exceeds the quantity studied in other populations and provided coverage of two-thirds of our taxonomy.

The distribution across types of questions reported is likely biased by our method of collection and the prompts provided to participants. Our instructions most likely elicited greater numbers of clinical information needs, and pregnant women have many additional medical and logistical needs that were not captured in our study. Given the time-dependent nature of the latter types of needs, participants may have addressed them through telephone or online messaging communications rather than recording them in their journal for review at a later date. Participants were advised that questions were being collected only for research purposes. They were encouraged to use their notebooks to remind them of questions for their physicians, but explicitly instructed that their physicians would not review their questions unless asked to by the participant. The limited racial diversity of the participants providing questions in our study may also limit the generalizability of our content distribution to populations with greater racial heterogeneity.

We achieved only moderate inter-rater reliability of the taxonomy when using the subcategory codes. As compared to other studies categorizing medical content, our coding reliability is stronger, especially considering the limited training our raters received. For example, in a study of primary care doctors evaluating clinical questions from other primary care and family doctors, they achieved 55% overall agreement with $\kappa = 0.53^{35}$. Further, our substantial agreement using only primary categories suggests that these categories are well-defined and mutually exclusive. We expect that with additional training and experience using our taxonomy, inter-rater agreement will increase substantially.

Conclusions

In this study, we have proposed and evaluated one of the first comprehensive taxonomies of consumer health information needs and applied it to describe the information needs of women carrying high-risk pregnancies. Pregnant women with prenatally-diagnosed congenital anomalies or pregnancy complications have significant and diverse information needs, and the observed needs could all be categorized using our simple but comprehensive taxonomy. The unmet needs identified in this population were largely related to clinical knowledge about the prognosis, diagnosis, management, and intervention plans for the fetus. Additional easily addressed logistical needs about healthcare facility policies were identified. We achieved moderate to substantial coding reliability with our proposed taxonomy. By modeling the spectrum of consumer health information needs, our taxonomy may guide content development for healthcare consumer information resources and improve automated systems to intelligently process and answer consumers’ questions.
Acknowledgments

This research project was funded by a grant from the Vanderbilt Children’s Hospital Children’s Development Fund. We acknowledge Mary Dabrowiak, Mariann Rimer, and staff at the Fetal Center at Vanderbilt for assistance with recruitment of research subjects for this study. We are indebted to the pregnant women who participated in this study.

References

Adaptation of a Published Risk Model to Point-of-care Clinical Decision Support Tailored to Local Workflow

Jeffrey L. Sobel, MD1, Craig C. Baker, BSN2, David Levy, MD1, Carol H. Cain, PhD2
1The Permanente Medical Group, Oakland, CA; 2Kaiser Permanente Care Management Institute, Oakland, CA

Abstract
Electronic clinical decision support can bring newly published knowledge to the point of care. However, local organizational buy-in, support for team workflows, IT system ease of use and other sociotechnical factors are needed to promote adoption. We successfully implemented a multi-variate cardiac risk stratification model from another institution into ours. We recreated the model and integrated it into our workflow, accessing it from our EHR with patient-specific data and facilitating clinical documentation if the user accepts the model results. Our clinical leaders championed the change and led educational dissemination efforts. We describe the ad-hoc social and technical collaboration needed to build and deploy the tool. The tool complements a clinical initiative within a community of practice, and is correlated with appropriate use of nuclear imaging.

Introduction
Electronic clinical decision support (CDS) shows promise for hastening the use of the latest evidence in clinical practice, but factors in addition to the clinical content influence adoption, including integration with workflow, presentation, and ease of use1,2. The challenge of localization means that CDS built at one institution is often not directly transferable to another. Furthermore, the adoption of a tool depends on local champions, support, and change management, components of the sociotechnical approach to information systems1,3.

We report on the successful adaptation of a published multi-variate cardiac risk model, the Cleveland Clinic treadmill mortality risk calculator4 to multiple medical centers in Kaiser Permanente (KP) Northern California. KP Northern California is an integrated healthcare delivery system providing comprehensive inpatient and outpatient care for more than 3.6 million members. KP physicians practice as part of a Permanente Medical Group, with peer-to-peer collaboration through Chiefs Groups. Each KP medical center has Chiefs of various areas, who have responsibility for oversight and improvement of clinical practice within their area. Most areas are clinical specialties, though there are Chiefs Groups for areas like Technology as well.

The Chiefs of Cardiology identified the need to better assess cardiac risk in patients during treadmill testing so that nuclear imaging tests, with their balance of harms and benefits, could be used more appropriately. They reached agreement about the best evidence-based clinical model, collaborated with KP’s Technology Chiefs and KP’s Clinical Library (CL) to create CDS that fit into daily workflow, and led integration and use by clinicians. We describe the sociotechnical steps we took to assure implementation, and report on use and associated outcomes.

Methods
Clinical need and clinical content. KP Chiefs Groups meet together regularly several times a year, led by a designated Chair. These peer groups act as change agents for general KP-wide initiatives, and also elevate specialty-specific concerns in order to draw upon KP-wide resources. The approximately 150 cardiologists in KP Northern California are represented through the Cardiology Chiefs. The Cardiology Chiefs reviewed region-wide utilization and determined that there was unwarranted variation in referrals to nuclear imaging after treadmill testing, that a common evaluation algorithm (the Duke TM score5) categorized low-risk patients as intermediate risk, and that nuclear imaging test results were not always needed to make the treatment decision. Nuclear imaging tests carry a risk of patient harm through unnecessary exposure to radiation. In addition, mis-categorizing patients at low risk of coronary artery disease can lead to further invasive evaluations or unwarranted medical therapy, where the risks outweigh the benefits. High-risk patients do often benefit from invasive procedures (e.g., cardiac catheterization), but low-risk patients can and should be effectively treated through medical management.
With the 2007 publication of a novel treadmill risk model, the Chair of the Cardiology Chiefs identified an opportunity to improve and align KP’s evaluation algorithm so that more low risk patients could be appropriately identified without need for nuclear imaging. The algorithm gives mortality prognosis to patient and providers, taking into account functional capacity and other proven parameters including heart rate recovery. The algorithm is appropriate for patients with a normal baseline ECG, no known coronary disease, cardiac or renal transplantation, pacemaker or defibrillator placement, or end-stage renal disease. This algorithm had improved prognostic accuracy, allowing for appropriate conservative management. The algorithm could be encapsulated into a CDS model with the data available at the time of the treadmill test. Furthermore, it had been tested and validated in a population that included KP patients. Before implementation, the algorithm was further evaluated on a historical sample of KP patients. The algorithm is not meant to be a replacement for clinical judgment, and predicts mortality but not coronary artery disease.

**Technology infrastructure.** KP uses an electronic health record (EHR) by Epic Systems with integrated CDS. We also have robust CDS that interfaces with the EHR but lives outside it, allowing us greater flexibility in programming logic and maintenance. Clinical calculators comprise one such category of CDS, hosted through KP’s web-based Clinical Library. We extracted patient-specific data from the EHR and replicated the calculations reported in the literature using Javascript. In testing, we found that our initial results did not match those reported in the literature. Subsequently, we contacted the original researchers at Cleveland Clinic. They assisted us in translating among programming languages and more accurately matching our model parameters to their model; limitations of paper journal publication prevented reporting of the entire model. Additional testing was performed to confirm model accuracy. The cardiac risk calculator is available to all KP clinicians through the intranet at KP facilities, individual log-on to CL, or a KP-specific Toolbar on the top of every EHR screen. The Toolbar view is specific to the user’s clinical role (in this case, the Treadmill department) and integrates calculator access into EHR workflow.

**Collaboration and Implementation.** An ad-hoc group of clinicians and technologists collaborated to build the calculator. The Cardiology Chiefs directed the clinical algorithm selection and the target users, as well as the form of the CDS. They elected to implement it as a consultation tool to complement clinical judgment rather than, for example, a required step of the workflow. These decisions led to conversations with different technology groups about CDS possibilities, including KP’s EHR development team, KP’s Toolbar team, and KP’s CL team. The calculator was tested with Cardiologists for ease of use, and functionality was built to show clinical definitions upon hovering over terms, and to generate a model report in a format that meets clinical documentation needs. The documentation was designed so the user could enter it into the patient chart with one click. The CDS calculator was developed over approximately 6 months. The algorithm endorsed by the Cardiology Chiefs and the calculator were publicized through regular meetings of the Chiefs Group and in-person conferences. The Cardiology Chiefs and other subject matter experts then carried the message to their own medical centers through verbal and written communication and specific in-person tutorials. The Cardiology Chiefs regularly review data on imaging use. Model use is tracked using Webtrends.

**Results**

The risk calculator became available to all KP users through Clinical Library in early 2010 (Figure 1), and is featured on the KP-specific Toolbar within the EHR for clinicians in the Treadmill department. It takes as input patient characteristics of age, sex, history of angina, diabetes, hypertension, smoking, treadmill test results of total METs, and ECG results. The age and sex can be pre-populated from the EHR, and the other fields are entered by the physician or clinician during the patient’s treadmill test. The model calculates 3, 5 and 10-year survival. The calculator generates a summary report which is shared with the patient and easily copied to their electronic chart as well. The workflow efficiency of an automatically generated report makes clinical documentation faster, more efficient, and more complete in contrast to manual transcription methods.
From 2010 to 2014, the calculator was accessed approximately 42,800 times, over 500 times a month (Figure 2). We count a visit as complete after 15 minutes of inactivity; if the page was accessed multiple times within 15 minutes, it is counted as one visit. We did not rigorously count the total number of eligible patients for the calculator, in part because, like all decision support, we would not necessarily expect it to be consulted for all eligible patients. A rough sense of the eligible population is the total number of outpatient treadmill referrals per month, approximately 3500 referrals; this referral volume includes patients who would be excluded from the algorithm, and does not include inpatient referrals. The Cardiology Chiefs’ communication, even before go-live, and the ease of access contributed to a large volume of views, achieving 500 views per month after the first three months. The calculator complements improvement initiatives to deliver patient-specific precision medicine, reduce unnecessary and potentially harmful testing, and spread best practices throughout KP. From 2006 to 2011, use of nuclear myocardial perfusion imaging declined by 51%8, after which imaging use remained flat. Over that same interval, KP maintained nation-leading outcomes in cardiac health9. The Cardiology Chiefs continue to receive and monitor imaging use data as a component of their quarterly dashboards.

---

**Figure 1.** Treadmill calculator on KP Clinical Library.

This calculator is designed to work with Internet Explorer only; other browsers may give erroneous results.

**Treadmill Mortality Risk Calculator**

*Hover over the text to the left of each data input field for a detailed description.*

<table>
<thead>
<tr>
<th>Age</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
</tr>
<tr>
<td>Typical Angina</td>
<td>No</td>
</tr>
<tr>
<td>Being Treated for Diabetes</td>
<td>No</td>
</tr>
<tr>
<td>Current or Recent Cigarette Smoker</td>
<td>Yes</td>
</tr>
<tr>
<td>Hypertension</td>
<td>No</td>
</tr>
<tr>
<td>Total METs Achieved</td>
<td>4</td>
</tr>
<tr>
<td>ST-Segment Depression</td>
<td>2</td>
</tr>
<tr>
<td>Test-Induced Angina</td>
<td>No</td>
</tr>
<tr>
<td>Abnormal Heart Rate Recovery</td>
<td>Yes</td>
</tr>
<tr>
<td>Frequent Ventricular Ectopy During Recovery</td>
<td>No</td>
</tr>
</tbody>
</table>

**Calculate Risk**

3-year survival = 96 %
5-year survival = 83 %
10-year survival = 64 %

*Based on this risk calculator, the patient is classified as high risk (3 year risk of death of >= 9%). Clinical correlation recommended but per ACC guidelines coronary angiography should be considered (if patient is a candidate).*

---

* This calculator is based on the study by: Lauer MS, Pothier CE, Magid DJ, Smith SS, Kattan MW. An Externally Validated Model for Predicting Long-Term Survival Exercise Stress Treadmill Testing in Patients with Suspected Coronary Artery Disease and a Normal Electrocardiogram. *Annals of Internal Medicine* 2007;147:821-829.

This calculator was developed with the permission and the generous assistance of the Cleveland Clinic.
Discussion

Electronic clinical decision support can be effective at bringing knowledge to the point of care, but several factors influence its success, including whether the CDS supports a broader change initiative. Our sociotechnical experience of bringing a published model into our institution reveals several lessons.

People. Identifying, building, and deploying a model of care required ad-hoc collaboration among many groups, even in an organization that has many established structures for collaboration. This effort included collaboration...
between multiple clinical and technology teams. KP’s Cardiology Chiefs were a critical driving force in leading this effort. The Chair of Chiefs provided the intellectual and clinical case for change. Indeed, the momentum for alignment around one algorithm preceded and fueled the creation of an electronic tool, especially as the discussions around tool creation took many months. The tool was a solution to the desire to make it easy for practicing clinicians to align with the recommended algorithm; the algorithm, while considered better than alternatives, primarily served to ensure consistency of care and reduce harm. The Cardiology Chiefs determined that the most effective target audience was cardiologists rather than primary care physicians, due to the volume of candidate patients each specialty was likely to see. They directed how the model would fit into clinical workflow, how users were likely to access the system, and the simplicity required in the user interface to the model. They also directed how the interface should yield further information about the model details in order to increase confidence in the model. Most importantly, they had the trust of their peers and understood their roles as champions of change. The tool became available in the midst of a campaign for change, and joined the arsenal of methods to create and sustain recommended referral criteria.

Replicating a published model was not completely straightforward. In re-building their model, we contacted the original researchers for their model parameters as well as technical mathematical implementation details that differed between programming languages. We sincerely thank the researchers for their generosity in sharing their model and their time.

Technology. We chose to replicate the model despite the researchers providing a publicly accessible web-based version. Before our tool was ready, the Cardiology Chiefs provided the external URL in their presentations to align care for nuclear imaging. We have found that externally hosted models are not reliably accessible, and indeed the original URL of the published model is no longer valid. A replicated model also allows us to create tighter data interfaces to our EHR. The technical replication of a model is not entirely straightforward. Our EHR does not have an easy interface to build, modify, customize or maintain an algorithmic calculator, so we decided to build it in CL, which hosts several KP-built calculators. Interoperability between the EHR and external CDS is a challenge. We initially built screen scraping functionality to extract patient age and sex from the EHR; of late we have been able to use Epic Systems’ ClinKB function to extract patient-specific data to our new calculators. We find that some data elements cannot be pulled discretely from the patient record, and had to provide the ability for users to enter the data quickly and easily into our calculator. By building CDS outside the EHR, we were able to add functionality that improved clinical adoption, including explanations of rationale and tailored clinical documentation.

Process. The collaborative designers of the tool followed principles associated with good clinical decision support, including ease of use, workflow integration, meeting a user need, and getting feedback. User-entered data is not written back into the EHR from the calculator. We automatically copy calculator results in summary form to the computer desktop clipboard, which users can then paste into a patient note. We believe this dramatically increased adoption because the function was easy to use and made it quicker to document the encounter, overcoming the inherent inertia associated with a change (Figure 4). We find that performing calculations outside of the EHR has several benefits. We have seen the calculator used as a simulation and patient education tool, where users enter hypothetical parameters to see the results. For example, the calculator can show patients the effect of quitting smoking on their mortality risk, an often powerful addition to a smoking cessation conversation. Having the calculator outside the EHR also allows clinicians to exercise clinical judgment about when the model applies.

The Chiefs Group endorsed and created awareness of the tool, in concert with the changes associated with appropriate imaging use. This involved duplicative communication: presentation at regular bi-monthly meetings, presentation at KP’s annual cardiology conference, and local peer-to-peer teaching. It also relied on clinicians’ pre-existing knowledge of where such tools are found: KP’s Clinical Library and the KP-developed toolbar within our EHR. The Chiefs Group regularly reviews quality outcomes including the appropriate use of potentially harmful imaging modalities. This pre-existing organizational structure and process allowed us to achieve steady-state use within months, made organizational deployment relatively easy, and is reusable for CDS addressing other topics. Intangible, but critical, is a physician leadership culture that supports patient-centered decisions, continuous improvement, reduction in unnecessary interventions, quality, and safety.
Limitations. The deployment of our calculator was designed to complement a broader clinical initiative, so it is difficult to isolate the effect of the tool itself on the observed outcomes. Nuclear imaging use has been on the decline throughout KP since 2006, and the calculator is one of several tools to support clinical judgment. This aligns with published evidence that multi-modal interventions are more effective than CDS alone at changing clinical practice10. The literature supports our hypothesis that peer-based communication is effective in changing practice, but we did not measure outcomes, such as intent-to-change self-assessments, associated with these communications.

Conclusion

KP is a large integrated delivery system with robust communities of medical sub-specialties, and established channels for delivering CDS. The replication and uptake of a new calculator was markedly easier because we did not have to create sociotechnical mechanisms specifically for one CDS tool. We find the adage, “If you build it, they will come” is true with qualifications. Building a tool, even a published one, is not entirely straightforward. Vigorous communication through peer communities, support of workflow, and ease of use are critical ensuring that CDS affects patient outcomes.
References


6. www.epic.com


Data-driven Temporal Prediction of Surgical Site Infection

Cristina Soguero-Ruiz MSc, Fei Wang M.E. PhD, Robert Jenssen MSc PhD, Knut Magne Augestad MD PhD, José-Luis Rojo Álvarez MSc PhD, Inmaculada Mora Jiménez MSc PhD, Rolv-Ole Lindsetmo MD PhD, Stein Olav Skrøvseth MSc PhD

1Department of Signal Theory and Communications, Telematics and Computing, Rey Juan Carlos University, Madrid, Spain; 2Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA; 3Department of Physics and Technology, University of Tromsø – The Arctic University of Norway, Tromsø, Norway; 4Department of Clinical Medicine, and 5Department of Mathematics and Statistics, University of Tromsø – The Arctic University of Norway, Tromsø, Norway; 6Norwegian Centre for Integrated Care and Telemedicine, and 7Department of Gastrointestinal Surgery, University Hospital of North Norway, Tromsø, Norway; 8Department of Surgery, Hammerfest Hospital, Hammerfest, Norway;

Abstract

Analysis of data from Electronic Health Records (EHR) presents unique challenges, in particular regarding non-uniform temporal resolution of longitudinal variables. A considerable amount of patient information is available in the EHR - including blood tests that are performed routinely during inpatient follow-up. These data are useful for the design of advanced machine learning-based methods and prediction models. Using a matched cohort of patients undergoing gastrointestinal surgery (101 cases and 904 controls), we built a prediction model for post-operative surgical site infections (SSIs) using Gaussian process (GP) regression, time warping and imputation methods to manage the sparsity of the data source, and support vector machines for classification. For most blood tests, wider confidence intervals after imputation were obtained in patients with SSI. Predictive performance with individual blood tests was maintained or improved by joint model prediction, and non-linear classifiers performed consistently better than linear models.

Introduction

When using observational data from secondary sources such as the Electronic Health Record (EHR) one needs to take into account that the information is rarely recorded in a systematic way. Indeed, the data are often sparse, and gathered at a clinician’s discretion. For example, blood tests are taken at a mixture of predefined stages in a patient pathway and clinically driven sampling. Thus, if predictive analytics relies on regularly sampled data, imputation methods need to be employed such that regular sampling is simulated. However, in the case of very irregular sampling a classical imputation approach may not be sufficient. In this paper we study prediction models for real time evaluation of patients admitted for gastrointestinal surgery with respect to surgical site infections (SSI) post-operatively.

SSIs are among the most common hospital-acquired infections. In fact, they represent up to 30% of all hospital acquired infections. 1,2 SSIs are associated with considerable morbidity and mortality. A mortality rate of 3%, prolonged stay up to 10 days and a significant decrease in quality of life, are reported. Similarly, readmissions related to SSIs are associated with a considerable increase in healthcare cost, up to 27,000 USD per readmission. 3 This persistent in-hospital morbidity is particularly associated with surgery for colorectal cancer. 4,6

The American College of Surgeons Surgical Quality Improvement Program (ACS-NSQIP) and The Centers for Disease Control and Prevention divide SSI into three subtypes based on the anatomical location of the infection, i.e. superficial, deep incisional and organ space. 4 Superficial infections can usually be cured with oral antibiotics and surgical debridement. In contrast, deep and organ space SSI require intravenous antibiotics, percutaneous drainage and laparotomies.

The patient specific risk factors for SSI are well documented and reported. A recent study by Lawson et al.4 identified open surgery, ulcerative colitis, older age, overweight, smoking, disseminated cancer and prolonged operation time as factors contributing to an increased risk of SSI. However, they found that different risk factors were associated with

1164
superficial and deep SSI. High Body Mass Index (BMI) and revision of an osteomy were associated with superficial SSI, whereas prolonged operation time and perioperative transfusions were associated with organ space SSI.\textsuperscript{4,6}

Using blood test results as predictive features in a data-driven decision support system is useful since these are performed relatively often with little burden to the patient. Therefore it is possible to estimate the expected information content of a blood test at stages in a patient trajectory.\textsuperscript{7} However, combining different tests performed at different stages in the trajectory, a necessity when observational data are used, presents challenges that we address here. The information from tests may be further combined with other data such as textual features that are predictive of complications.\textsuperscript{8}

For the purpose of this manuscript we denote the sparsity of the clinical data as missing data. Missing data percentages are even larger for some studies such as clinical laboratory measurements or biomarkers. Despite of the efforts made to develop statistical methods for handling missing data, there is no global best approach because they inevitably depend on stated assumptions.

In this work we present methods for predictive modeling in a context of features that have strongly irregular sampling patterns. We analyze different smoothing and interpolation/imputation techniques and different input spaces to predict SSI using blood tests. Finally we look at linear and non-linear classifiers to do the predictive modeling. Figure 1 shows an overview of the data-driven decision support system used in this work for SSI prediction.

\textbf{Methods}

We extracted a cohort of patients based on relevant International Classification of Diseases (ICD10) or NOMESCO Classification of Surgical Procedures (NCSP) codes related to severe post-operative complications, and in particular to surgical site infections, from the EHR of the Department of Gastrointestinal Surgery at the University Hospital of North Norway. The selection of codes was guided by input from clinicians at the hospital. The cohort identified as control was matched with patients that did not have any of these codes but were otherwise similar in terms of which blood tests were performed. Additionally, a text search was performed to ensure that the controls did not have the word “infection” in any of their post-operative text documents. This resulted in a cohort of 101 cases and 904 matched controls. Patients with codes indicating superficial infections were excluded. A set of 10 different types of blood tests were defined as clinically relevant and extracted for all patients from their EHRs. All tests were not available every day, which results in a high percentage of missing values when analyzing data on that scale, yielding to a non-uniform time sampling description for each patient (Fig. 2). The data matrix is hence sparse over lab tests and time, therefore constituting a challenging data set to work on. A method denoted bootstrap nonparametric resampling\textsuperscript{9,10} was designed to statistically describe the influence of imputation. Thus, the population mean and corresponding 95\%
Table 1: Demographic characteristics of the patient groups.

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Controls</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (%)</td>
<td>477 (47.4)</td>
<td>441 (48.7)</td>
<td>36 (35.6)</td>
</tr>
<tr>
<td>Age [Mean±SD]</td>
<td>57.0 ± 20.7</td>
<td>56.9 ± 21.2</td>
<td>57.4 ± 15.2</td>
</tr>
</tbody>
</table>

Figure 2: Available laboratory test measurements for one control (a) and another infected (b) patient. The $y$-axis shows the available tests for a patient, with no specific order, as each patient can have different number of tests. The $x$-axis represents the day when each test was recorded, being $t = 0$ the day when the first test was recorded. Vertical black line indicates the surgery day, whereas red line indicates the infection day.

CI was computed on a daily basis for each test, obtaining an averaged trend.

The data represent a diverse group of patients undergoing gastrointestinal surgery such that results can generalize across this group. The basic demographics of the cohort are given in Table 1.

**Feature extraction for sparse clinical data.** Working with complete datasets is the standard scenario for most statistical and machine learning methods. In the literature, there are works that simply omit patients with any missing data. However, discarding patients with missing data may lead to incorrect assessments or prognostics. To avoid this situation, different methods have been proposed to deal with observations at non regular sampling. These methods can be categorized into: (1) smoothing or interpolation techniques; (2) spectral analysis tools such as wavelets or Lomb-Sargle Periodogram; and (3) kernel methods.

Regarding interpolation methods, the well-known Last Observation Carried Forward (LOCF) scheme imputes the last non-missing value for the following missing values. Some works support that LOCF should not be considered as the primary approach to the treatment of missing data. Alternatively, Lasko et al. suggest using Gaussian Process (GP) followed by a warped function, and we follow this approach in this paper. The warped function is intended to adjust for the fact that rapid changes in temporal variables in connection with active treatment is often followed by long periods of apparent stability leading to highly nonstationary processes. The time warping function can be constructed as

$$d' = d^{1/\alpha} + \beta$$  

where $d$ is the original distance between two adjacent observations, and $\alpha$ and $\beta$ are free parameters to be tuned. This function converts non-stationary clinical data into a stationary process which allows the use of a GP to deal with sparsity.

A random process $f(t)$ is a Gaussian process if, for any finite set of values of $t_1, t_2, \ldots, t_k$, the variables of the corresponding random vector $\mathbf{f} = f(t_1), f(t_2), \ldots, f(t_k)$ are jointly normal (Gaussian). Element $K_{ij}$ of the covariance
matrix $K$ of $f$ is $k[f(t_i), f(t_j)]$ where $k[·, ·]$ is a covariance (kernel) function, such as the radial basis function, or the squared exponential function. Using Bayes theorem, the posterior density function for an (unseen) random variable $f_* = f(t_*)$ conditioned on the observed $f$ becomes
\[
P(f_*|f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(f_* - \hat{f})^2}{2\sigma^2} \right],
\]
where the posterior mean value is given by $\hat{f} = k^T K^{-1} f$; and the posterior variance is $\hat{\sigma}^2 = \kappa - k^T K^{-1} k$. In this expression, element $i$ of the vector $k$ is $k[f(t_i), f(t_i)]$, $i = 1, \ldots, k$, and $\kappa = k[f(t_i), f(t_i)]$. In Gaussian process regression, $\hat{f}$ is used as the estimate, or prediction, of $f_*$, while $\hat{\sigma}^2$ provides the level of confidence in the prediction. Thus, we use this approach in this manuscript to deal with sparse data.

**Prediction Analytics.** The SSI prediction is performed from a data-driven approach based on machine learning techniques. These techniques learn the underlying predictive model from a set of $d$-dimensional examples known as training set, where $d$ corresponds to the number of features which are supposed to be relevant to the predictive task. The generalization of the model is estimated using an independent set of examples known as test set. Among the plethora of machine learning techniques proposed in the literature, Support Vector Machines (SVM) have shown to provide good generalization capabilities, and they have been considered in this work. We next briefly describe linear and non-linear SVM classifiers as well as the feature selection (FS) methods used in this work.

**Linear and non-linear SVM.** The data model for a linear classifier is given by $y = \langle x, w \rangle + b$, where $x$ is the feature vector, $w$ is the weight vector of the linear model, $b$ is the bias term, $y$ is the classification output and $\langle ·, · \rangle$ denotes the inner product. We focus on the Support Vector Machine (SVM) classifier, with a regularization term such that model complexity is controlled, and the upper bound of the generalization error is minimized. These theoretical properties make the SVM an attractive approach for predictive modeling.

Denote $D = \{x_i, y_i\}_{i=1}^n$ as a binary labeled training set, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. The SVM classifier seeks the separating hyperplane with the largest margin between the two classes. The hyperplane optimally separating data is defined from a subset of training data named support vectors (SV), and it is obtained by minimizing $\|w\|^2$, as well as the classification losses in terms of a set of slack variables $\{\xi_i\}_{i=1}^n$. Considering the $\nu$-SVM introduced by Schölkopf et al. and a potential non-linear mapping $\phi(·)$, the $\nu$-SVM classifier solves
\[
\min_{w, (\xi_i), \nu, \rho} \left\{ \frac{1}{2} \|w\|^2 + \nu \rho + \frac{1}{n} \sum_{i=1}^n \xi_i \right\}
\]
subject to:
\[
y_i(\langle \phi(x_i), w \rangle + b) \geq \rho - \xi_i, \quad \rho \geq 0, \quad \xi_i \geq 0 \quad \forall i = 1, \ldots, n.
\]
The variable $\rho$ adds another degree of freedom to the margin, and the margin size linearly increases with $\rho$. The parameter $\nu \in [0, 1]$ acts as an upper bound on the fraction of margin errors, and it is also a lower bound on the fraction of SVs. An appropriate choice of non-linear mapping $\phi$ guarantees that the transformed input vectors are more likely to be linearly separable in the (higher dimensional) feature space.

The primal problem in Eq. (3) can be solved using its dual formulation, yielding $w = \sum_{i=1}^n y_i \alpha_i \phi(x_i)$, where $\alpha_i$ are Lagrange multipliers corresponding to constraints in Eqs. (3–4). Thus, the decision function for any test vector $x_*$ is given by
\[
f(x_*) = \sum_{i=1}^n y_i \alpha_i K(x_i, x_*) + b
\]
In order to predict the label of $x_*$, the sign of $f(x_*)$ is used. The so-called SV are those training samples $x_i$ with corresponding Lagrange multipliers $\alpha_i \neq 0$. The bias term $b$ is calculated by using the unbounded Lagrange multipliers as $b = 1/k \sum_{i=1}^k (y_i - \langle \phi(x_i), w \rangle)$, where $k$ is the number of non-null and unbounded Lagrange multipliers.
The use of Mercer kernels allows to handle the non-linear algorithm implementations as $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. In this work, we use two well-known Mercer kernels: the linear kernel $K(x, z) = \langle x, z \rangle$, and the Radial Basis Function (RBF) kernel $K(x, z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$, where $\sigma$ is the width parameter, to be tuned together with $\nu$ free parameter.

**SVM Feature Selection.** Feature Selection (FS) strategies have been widely studied in the machine learning literature. The purpose of FS is to choose a subset of features that are relevant for classification or regression tasks, while at the same time maintain or improve the performance of the learning method in comparison of using the whole set of available features. Regarding FS for SVM linear classifiers, the Recursive Feature Elimination (RFE) method has been shown to compare very favorably to many of the classical FS methods. Improved versions of SVM FS methods that included non-linear kernel functions have been described.

**Results**

In this section, we evaluate the capabilities of different ways to deal with sparse data and show the effects on performance results. Furthermore, linear and non-linear classifiers are benchmarked to predict SSI when using data from different laboratory tests obtained from the EHR. Firstly, each laboratory test was used separately to predict SSI using linear and non-linear classifiers after dealing with sparse data. Secondly, we analyzed the use of multiple blood tests to check the impact of combining them as well as the temporal-feature relative importance.

Our database was imbalanced, with 101 and 904 cases in the positive and negative classes, respectively. This is a common situation for clinical databases, where different number of patients are assigned to each class. Though previous studies have demonstrated that balanced classes in the training set often improve the overall classification performance, we used an undersampling strategy (discarding samples from the majority class), such that the training set was built by enforcing balanced classes. In order to represent correctly the population, we selected a different number $S$ of subsets of the negative class with 101 samples in each, and computed classification performances in terms of the mean and the standard deviation of the results for each subset.

We used a cross-validation strategy to ensure the generalizability of the prediction analytics. First, we balanced the classes and then we split the data into training and tests subsets (80%-20%). A leave-one-out (LOO) cross-validation was carried out on the training subset of the balanced set for selecting the classifier free parameters. Thus, the SVM classifier was retrained $R$ times, where $R$ is the number of cases per class for balanced classes ($R = 101$ in this work). In this work, accuracy was considered as performance measurement for free parameter tuning.

**Effect of the imputation methods on performance.** Two different strategies, namely, LOCF and warped-GP, were considered to deal with the extreme sparsity present in the input space as given by different tests measured in a patient at different days.

*Last observation carried forward (LOCF).* The last observed non-missing value was used to fill in the missing values into a regular time sampling grid with a daily time basis, i.e., if we have a missing value, we consider instead the previous value if it exists. A nonparametric resampling method to represent the averaged trend was applied to statistically describe the influence of imputation. See two examples in Fig. 3 (a) for C-Reactive Protein (CRP) and Fig. 3 (b) Potassium tests. It is well known that CRP is a good predictor for complications after colorectal surgery. We note that our pattern of CRP levels following surgery is consistent with that observed by Singh et al.

For most blood tests, wider confidence interval (CI) after LOCF imputation were obtained for patients with SSI. Specifically, the data recorded at the day of surgery are highly noisy, as it can be seen in Fig. 3. For this reason, we excluded these values from our analysis, and we focused only on pre-operative and post-operative periods.

*Warped function and GP.* Using the time warped function Eq. (1), for each test we selected values of $\alpha$ and $\beta$ parameters which maximize the accuracy of the predictive system. For this purpose, we used a grid search over values $\alpha \in [1, 10]$ and $\beta \in [0, 100]$. A LOO strategy was considered to ensure generalizability. The use of GP regression allows us to transform a set of finite measurements contained in the EHR from each blood tests into a continuous longitudinal function. In this way, missing values are inferred, allowing pre-operative and post-operative feature extraction.
Prediction of SSI. Table 2 shows the pre-operative and post-operative classification performance in terms of accuracy (mean and 95% CI) for each blood test individually when considering a LOCF strategy and warped function with GP methodology. Pre-operative stage was defined as four days before surgery, and four days immediately after surgery were considered in the post-operative stage (i.e., $d = 4$). We considered linear and non-linear SVM classifiers for the prediction of SSI, and we benchmarked results with a simpler logistic regression classifier. Results suggest the presence of strong non-linear relationship among input features for the analyzed tests, as given by consistently achieving the best performances when a non-linear SVM was considered. Note also that the post-operative predictive power is in general higher than pre-operative, which is to be expected.

Table 2 also shows that performance depend on the method used to deal with sparsity. In general, the combination of warped function and GP improved results, however, it can be seen that for some tests LOCF is better.

Feature selection. The results in Table 2 suggest that a non-linear classifier provides a better prediction of SSI. Taking that into account, we obtained the accuracy using a non-linear SVM classifier both for pre-operative and post-operative stages. First, we considered all blood tests together, i.e., $d = 40$ (first row in Table 4) and then we applied the FS method denoted as RBF RFE (second row in Table 3). Table 3 shows the mean and 95% CI accuracy when using both schemes. Comparison of Table 2 and Table 3 shows that the model built with all tests provide higher accuracy. Note also that a similar or tending to higher accuracy is obtained with the FS method, so it is appropriate for addressing the interpretation of the relevance and meaning of the input space.

Figure 4 summarizes the results of FS with non-linear SVM (with RBF kernel) in terms of relevance of blood tests. Towards that end, we calculated how many times every feature is selected (frequency of relevance), separately for the pre-operative and post-operative stages. From these values, a relevance index for each blood test is obtained as the normalization of the cumulative frequency of relevance by number of features per day ($d = 4$) times the number of subsets ($S = 5$). Note that a comparison with baseline level is remarkable for all tests (excepts sodium), indicating the relevance of the intra-patient pre-operative levels on each test. In general terms, thrombocytes reached the highest prediction information, together with ALP, CRP, albumin, creatinine and leukocytes, most of them being consistent with previous results. Although less relevant in the pre-operative state, the other tests (potassium, ALAT, and hemoglobin) also included highly relevant information in the post-operative state.
Table 2: Pre-operative and post-operative prediction results in terms of accuracy (mean and 95% CI) for each test individually and different classifiers: Logistic regression (first row), linear SVM (second row), and non-linear SVM (third row). The best accuracy values for pre-operative and post-operative are shown in bold.

<table>
<thead>
<tr>
<th>Lab test</th>
<th>LOCF Pre-operative</th>
<th>LOCF Post-operative</th>
<th>Warped-GP Pre-operative</th>
<th>Warped-GP Post-operative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemoglobin</td>
<td>0.48 [0.43,0.53]</td>
<td>0.47 [0.44,0.75]</td>
<td>0.60 [0.54,0.64]</td>
<td>0.60 [0.54,0.64]</td>
</tr>
<tr>
<td></td>
<td>0.58 [0.50,0.69]</td>
<td>0.62 [0.51,0.69]</td>
<td>0.52 [0.40,0.62]</td>
<td>0.55 [0.46,0.63]</td>
</tr>
<tr>
<td></td>
<td>0.70 [0.56,0.84]</td>
<td><strong>0.89 [0.77,0.95]</strong></td>
<td><strong>0.71 [0.64,0.81]</strong></td>
<td>0.79 [0.65,0.85]</td>
</tr>
<tr>
<td>Leucocytes</td>
<td>0.50 [0.43,0.56]</td>
<td>0.47 [0.43,0.51]</td>
<td>0.54 [0.48,0.59]</td>
<td>0.54 [0.48,0.59]</td>
</tr>
<tr>
<td></td>
<td>0.50 [0.38,0.59]</td>
<td>0.61 [0.50,0.71]</td>
<td>0.45 [0.30,0.55]</td>
<td>0.53 [0.44,0.65]</td>
</tr>
<tr>
<td></td>
<td><strong>0.75 [0.62,0.85]</strong></td>
<td>0.77 [0.65,0.85]</td>
<td><strong>0.75 [0.61,0.87]</strong></td>
<td><strong>0.81 [0.73,0.93]</strong></td>
</tr>
<tr>
<td>CRP</td>
<td>0.49 [0.44,0.55]</td>
<td>0.48 [0.44,0.54]</td>
<td>0.62 [0.51,0.73]</td>
<td>0.44 [0.41,0.50]</td>
</tr>
<tr>
<td></td>
<td>0.51 [0.43,0.60]</td>
<td>0.79 [0.71,0.87]</td>
<td>0.50 [0.39,0.67]</td>
<td>0.60 [0.47,0.71]</td>
</tr>
<tr>
<td></td>
<td>0.61 [0.52,0.69]</td>
<td><strong>0.90 [0.84,0.94]</strong></td>
<td><strong>0.79 [0.66,0.94]</strong></td>
<td>0.79 [0.67,0.88]</td>
</tr>
<tr>
<td>Potassium</td>
<td>0.48 [0.44,0.54]</td>
<td>0.47 [0.44,0.54]</td>
<td>0.52 [0.49,0.60]</td>
<td>0.48 [0.51,0.44]</td>
</tr>
<tr>
<td></td>
<td>0.58 [0.49,0.66]</td>
<td>0.64 [0.46,0.72]</td>
<td>0.59 [0.52,0.69]</td>
<td>0.53 [0.63,0.43]</td>
</tr>
<tr>
<td></td>
<td><strong>0.73 [0.60,0.84]</strong></td>
<td><strong>0.88 [0.77,0.95]</strong></td>
<td>0.66 [0.60,0.83]</td>
<td>0.74 [0.64,0.86]</td>
</tr>
<tr>
<td>Sodium</td>
<td>0.48 [0.44,0.54]</td>
<td>0.47 [0.44,0.54]</td>
<td>0.49 [0.45,0.57]</td>
<td>0.48 [0.42,0.53]</td>
</tr>
<tr>
<td></td>
<td>0.53 [0.43,0.68]</td>
<td>0.55 [0.34,0.73]</td>
<td>0.54 [0.42,0.70]</td>
<td>0.52 [0.46,0.58]</td>
</tr>
<tr>
<td></td>
<td>0.66 [0.56,0.74]</td>
<td><strong>0.76 [0.67,0.89]</strong></td>
<td><strong>0.71 [0.55,0.90]</strong></td>
<td>0.68 [0.63,0.79]</td>
</tr>
<tr>
<td>Creatinine</td>
<td>0.46 [0.40,0.53]</td>
<td>0.46 [0.44,0.50]</td>
<td>0.49 [0.47,0.57]</td>
<td>0.41 [0.34,0.45]</td>
</tr>
<tr>
<td></td>
<td>0.55 [0.46,0.62]</td>
<td>0.61 [0.44,0.67]</td>
<td>0.50 [0.36,0.59]</td>
<td>0.52 [0.38,0.64]</td>
</tr>
<tr>
<td></td>
<td><strong>0.79 [0.73,0.86]</strong></td>
<td>0.69 [0.56,0.82]</td>
<td>0.68 [0.55,0.74]</td>
<td><strong>0.75 [0.69,0.83]</strong></td>
</tr>
<tr>
<td>ALAT</td>
<td>0.50 [0.44,0.53]</td>
<td>0.49 [0.44,0.53]</td>
<td>0.57 [0.49,0.64]</td>
<td>0.54 [0.48,0.58]</td>
</tr>
<tr>
<td></td>
<td>0.61 [0.53,0.69]</td>
<td>0.54 [0.43,0.66]</td>
<td>0.63 [0.56,0.59]</td>
<td>0.49 [0.40,0.59]</td>
</tr>
<tr>
<td></td>
<td>0.69 [0.50,0.82]</td>
<td>0.61 [0.47,0.71]</td>
<td><strong>0.76 [0.63,0.88]</strong></td>
<td><strong>0.67 [0.63,0.75]</strong></td>
</tr>
<tr>
<td>Thrombocytes</td>
<td>0.57 [0.48,0.63]</td>
<td>0.56 [0.47,0.62]</td>
<td>0.57 [0.49,0.65]</td>
<td>0.57 [0.54,0.60]</td>
</tr>
<tr>
<td></td>
<td>0.56 [0.45,0.70]</td>
<td>0.66 [0.59,0.73]</td>
<td>0.61 [0.40,0.75]</td>
<td>0.49 [0.43,0.56]</td>
</tr>
<tr>
<td></td>
<td><strong>0.73 [0.62,0.83]</strong></td>
<td><strong>0.73 [0.66,0.89]</strong></td>
<td>0.65 [0.58,0.70]</td>
<td>0.68 [0.58,0.74]</td>
</tr>
<tr>
<td>Albumin</td>
<td>0.53 [0.41,0.65]</td>
<td>0.50 [0.41,0.64]</td>
<td>0.56 [0.52,0.60]</td>
<td>0.47 [0.42,0.50]</td>
</tr>
<tr>
<td></td>
<td>0.55 [0.40,0.66]</td>
<td>0.70 [0.44,0.84]</td>
<td>0.79 [0.55,0.92]</td>
<td>0.63 [0.54,0.69]</td>
</tr>
<tr>
<td></td>
<td>0.71 [0.48,0.89]</td>
<td>0.82 [0.69,0.93]</td>
<td><strong>0.91 [0.88,0.92]</strong></td>
<td><strong>0.83 [0.77,0.92]</strong></td>
</tr>
<tr>
<td>ALP</td>
<td>0.49 [0.38,0.54]</td>
<td>0.49 [0.41,0.53]</td>
<td>0.41 [0.36,0.54]</td>
<td>0.33 [0.31,0.36]</td>
</tr>
<tr>
<td></td>
<td>0.55 [0.43,0.67]</td>
<td>0.58 [0.53,0.65]</td>
<td>0.69 [0.64,0.75]</td>
<td>0.55 [0.44,0.71]</td>
</tr>
<tr>
<td></td>
<td><strong>0.69 [0.50,0.84]</strong></td>
<td>0.63 [0.47,0.76]</td>
<td><strong>0.69 [0.44,0.87]</strong></td>
<td><strong>0.74 [0.69,0.79]</strong></td>
</tr>
</tbody>
</table>

CRP: C-Reactive Protein; ALAT: Alanine aminotransferase; ALP: Alkaline phosphatase

Discussion

The results clearly demonstrate the utility of blood tests for predicting SSIs both pre- and post-operatively. These results will potentially be useful as part of a data-driven online clinical decision support system that can enable clinicians to improve post-surgical recovery rates. With proper warning necessary actions such as closer follow up and risk stratification can be performed.

We chose to generate the cohort in a way that may open the problem to being in a sense “too easy”. Since our emphasis was on methodology development, we chose to generate a cohort where the testing patterns were similar in the negative and positive classes. This aids the algorithmic development, but opens the possibility that the problem does not entirely reflect the clinical scenario since many in the negative class would not be suspected of having SSIs. Nevertheless, this does not invalidate the methodology development or the results. Indeed, if the envisioned decision support system is thought of as a warning system flagging patients at risk, the cohort reflects well the actual clinical setting.
Table 3: Joint pre-operative and post-operative accuracy (mean and 95% CI) with a RBF RFE FS method.

<table>
<thead>
<tr>
<th></th>
<th>LOCF</th>
<th>Warped-GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>All tests</td>
<td>0.81 [0.76,0.86]</td>
<td>0.89 [0.92,0.97]</td>
</tr>
<tr>
<td>FS</td>
<td>0.83 [0.67,0.90]</td>
<td>0.91 [0.90,0.92]</td>
</tr>
</tbody>
</table>

Figure 4: Pre-operative and post-operative relevance index for each blood test using LOCF (a) and Warped-GP (b).

Testing is often done at the discretion of the clinician, and often not driven by formulaic rules. This is part of the reason for the irregular sampling in the data, leading to the problem formulation in this paper. Thus, the testing pattern for patients in inpatient care itself may be an informative feature of post-operative complications independent of the test results. By generating the cohort by matching testing patterns, this information is largely lost and only the test results remain as the informative features.

In retrospective EHR studies there is inevitably the chance of a censoring effect where a test result informs the clinician of a possible complication and the clinician takes appropriate and successful action to avoid the complication. Then the pattern for complication will be present, but not the complication itself, which leads to effectively mislabeled data, known as confounding medical interventions. In our case this is unlikely to be a large issue since there is little information to act on to avoid a SSI such that most cases are likely to be correctly coded.

Using ICD10 and NCSP codes to phenotype a cohort there is a significant chance of miscoding leading to labeling errors. However, there is a far greater chance of false negatives (i.e., missing coding) than false positives. In this case, the positive class will be correct while the negative class may contain erroneous labels. When generating the cohort by matching we alleviate this since a minority of patients get SSIs, and additionally we check for the Norwegian equivalent of the word “infection” in the post-operative notes, which would almost surely appear if the patient actually got an SSI.

In the literature several approaches for reducing SSI have recently been described. Wick et al, used a multidisciplinary approach to reduce SSI and showed that formation of small groups of front line providers to address SSIs reduced SSIs by 33%. One of the most popular existing risk models for SSI is the National Nosocomial Infection Surveillance (NNIS) Basic SSI Risk Index. Also, recently, a logistic regression model for predicting SSIs was developed by van Walraven et al. A more data driven approach has been used by Gbegon et al. predicting SSIs in real time within 30 days of the operation. However, these studies rely on clinical data, demographic and other information but do not take blood tests into account. There exists validated risk assessment tools for post-operative complications, including the Surgical AGPAR Score and the POSSUM score. Both of them assess the immediate post-operative risk based...
on a number of variables. The American College of Surgeons’ NSQIP risk calculator was developed as a preoperative risk stratification tool.\textsuperscript{33}

In this context our work presents a path forward to combine clinical variables along with demographic and other data with test results that can be updated in real-time and provide a live assessment of a patient’s progression.

**Conclusion**

We have shown that our model has a potential for real time prediction and identification of patients at risk for developing SSI. This can give decision support to clinicians, and treatment plans can be adjusted taking into account the identified increased risk.

Appropriately adjusting the temporal structure of blood tests can dramatically improve the system accuracy. This can provide the basis for a future on-line system that alerts clinicians to patients at risk for complications, such that appropriate action can be taken. With early identification of these patients, improved clinical outcomes, reduced readmissions and cost savings are likely.

**References**


Ginkgo and Warfarin Interaction in a Large Veterans Administration Population

Gregory J. Stoddard¹, MS, Melissa Archer, PharmD², Laura Shane-McWhorter, PharmD², Bruce E. Bray, MD³, Doug F. Redd, MS³, Joshua Proulx, MS³, Qing Zeng-Treitler, PhD³,⁴

¹Department of Internal Medicine, University of Utah
²College of Pharmacy, University of Utah
³Department of Biomedical Informatics, University of Utah
⁴Veteran Affairs Salt Lake City Healthcare System

Abstract

Ginkgo biloba is a widely used herbal product that could potentially have a severe interaction with warfarin, which is the most frequently prescribed anticoagulant agent in North America. Literature, however, provides conflicting evidence on the presence and severity of the interaction. In this study, we developed text processing methods to extract the ginkgo usage and combined it with prescription data on warfarin from a very large clinical data repository. Our statistical analysis suggests that taking concurrently with warfarin, ginkgo does significantly increase patients’ risk of a bleeding adverse event (hazard ratio = 1.38, 95%CI: 1.20 to 1.58, p<.001). This study also is the first attempt of using a large medical record database to confirm a suspected herb-drug interaction.

Introduction

The National Center for Complementary and Alternative Medicine (NCCAM) is an agency in the U.S. Department of Health and Human Services dedicated to defining the usefulness and safety of complementary therapies through rigorous scientific research. Complementary and alternative medicine (CAM) is a term used to describe two different treatments. NCCAM defines “complementary therapies” as “non-mainstream” approaches used in combination with allopathic medicine, whereas “alternative therapies” are those that replace conventional medicine with non-mainstream approaches (1). Although CAM was previously divided into different categories, NCCAM states there are two basic subgroups – natural products, and mind and body practices. Natural products consist of botanicals (including herbs), vitamins and minerals, and probiotics. These natural products are commonly found in dietary supplements.

According to the National Health and Nutrition Examination Study (NHANES 2003-2006), approximately half of all Americans use supplements and spend $15 billion annually (2,3). Ginkgo biloba is one of the most purchased dietary supplements used in the United States, and is used to treat a variety of conditions such as memory deficits or dementia, intermittent claudication, tinnitus, and many other health concerns. One of the main concerns with ginkgo use is increased bleeding risk. Ginkgo may decrease platelet aggregation and many case reports have suggested increased bleeding risk, as verified by a systematic review (4). The increased bleeding risk posed by ginkgo may therefore be of great concern in persons taking anticoagulants such as warfarin.

Warfarin is one of the most frequently prescribed anticoagulant agents in North America. Warfarin works by blocking the effects of vitamin K, inhibiting the synthesis of clotting factors and preventing thromboembolic events. Despite its widespread use, warfarin therapy is associated with increased risk of hemorrhage and interacts with various medications, dietary supplements, and some foods. Patients taking warfarin are monitored closely for abnormal or increased bleeding and receive frequent blood testing for the prothrombin time international normalized ratio (INR) to ensure the warfarin dose is adequate yet safe. A high, out-of-range INR is often associated with warfarin drug-drug or drug-herb interactions and indicates increased risk of bleeding.

The interaction between ginkgo and warfarin has not been adequately studied in patients taking both products in combination. Since ginkgo combined with warfarin is a potentially severe interaction that may result in bleeding, we chose it as the first test case in a research study that evaluated the possibility of this occurrence in patients identified as being on both products. In this study, we mined a large national clinical data repository called VINCI to
investigate the ginkgo-warfarin interaction. With over 20 million unique patients and extensive medical records, we were able to identify thousands of patients who were using ginkgo and warfarin concurrently.

**Methods**

**Data Source**

We used existing electronic medical record data from the Veterans Administration (VA) Informatics and Computing Infrastructure (VINCI) database. Available to VA researchers, this database includes over 20 million unique patient electronic medical records from all VA hospitals and clinics in the United States, which are compiled using uniform coding of data elements. VINCI also includes a suite of research tools to facilitate analysis, such as natural language processing.

To identify concurrent usage of ginkgo and warfarin, we queried both structured and free text data in the VINCI database. We queried the clinical documents table for information containing the terms “ginkgo” and variants “ginkgo” and “ginko.” Of the matching documents approximately 50% used the terms ginkgo, 25% gingko, and 25% ginko. We also queried the filled prescriptions table for the term “warfarin” and its alternate brand names (Jantoven, Coumadin, Marevan, Lawarin, Waran, and Warfant). There were no occurrences of the alternate brand names, since they are not on the VA formulary, so future queries only used warfarin.

**Natural Language Processing (NLP)**

An NLP module was developed to further process the notes retrieved by the ginkgo query. We prioritized the NLP of ginkgo cases because little structured data are available for herbal supplement usage (Table 1).

For NLP development, we randomly selected 100 patients with notes containing any mention of ginkgo or one of its spelling variants (n=441) to create an annotated data set. Two reviewers developed a guideline to establish true positive cases and conducted chart review. The inter-reviewer agreement was calculated (Cohen’s kappa = 0.82).

Based on the manual review, we first crafted a set of processing rules to classify highly prevalent document templates (n=41). These processing rules identified positive occurrences of ginkgo in patient supplement lists recorded within the documents as well as negative occurrences of ginkgo in standard documents instructing the patient not to take ginkgo prior to an upcoming surgery. Then using the annotated documents that do not contain templates, we trained a support vector machine (SVM) model to classify the remaining notes not covered by the template rules. The SVM developed was conducted using the Waikato Environment for Knowledge Analysis (WIKAI) sequential minimal optimization (SMO) algorithm with the default parameters and bag-of-word features. The final NLP module first applies the template rules and then applies the SVM model.

To test the NLP module, we further annotated another 200 randomly selected notes retrieved by the ginkgo query and calculated the sensitivity and specificity of the NLP module. On the 200 randomly selected ginkgo notes, the NLP model reached a sensitivity of 97%, specificity of 87%, and F measure of 93%. Applying this NLP model to all ginkgo related notes (n=836,506), 600,107 documents and 132,061 patients were identified as positive.

The documentation of ginkgo usage often does not specify the start date or duration. The warfarin exposure was calculated using the VINCI pharmacy fill record. Co-administration was established when patients were exposed to both ginkgo and warfarin. Combining the NLP results with medical fill records, we found 54,139 combined use events in 9,862 distinct patients (Table 2).

<table>
<thead>
<tr>
<th>Ginkgo + warfarin</th>
<th># Patients (Structured Data)</th>
<th># Patients (Free Text Notes + Structured Data)</th>
<th># Overlapping Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9</td>
<td>9,862</td>
<td>8</td>
</tr>
</tbody>
</table>

**Sample Size**

The study sample consisted of all patients in the VINCI database that had at least one warfarin order during the study period (years 2008 to 2014), which provided a sample size of n=807,399 patients. Of these, n=11,003 also used ginkgo at least once concurrently with warfarin, and so composed the warfarin + ginkgo group (gingko group). The remaining n=796,396 formed the warfarin only group (non-gingko group).
Bleeding Events

Bleeding is the most frequent complication of warfarin therapy (5). A large body of evidence evaluating the safety of warfarin therapy is available. In clinical trials, bleeding events are classified as fatal, major, life-threatening, clinically significant, overt, or minor. ‘Major bleeding’ is the most common safety outcome cited in clinical trials but the definition varies across trials. According to the International Society on Thrombosis and Hemostasis, the definition of major bleeding should be based on objective criteria and only include events which are life-threatening, utilize major health-care resources, or result in death (5). A list of ICD-9 codes (Table 2) for bleeding events based on the above criteria have been identified and used in a number of clinical trials and analyses evaluating warfarin safety (5-8). We used this list to identify bleeding events in our patient population (Table 3).

Table 2. ICD-9-CM Codes for Major Bleeding Events

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>ICD-9 Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gastrointestinal Bleeding</td>
<td>530.82, 531.2, 531.4, 531.6, 532.2, 532.4, 532.6, 533.2, 533.4, 533.6, 534.2, 534.4, 534.6, 535.x1, 537.83, 562.02, 562.03, 562.12, 562.13, 569.3, 578.x</td>
</tr>
<tr>
<td>Intracranial Bleeding</td>
<td>430.x, 431.x, 432.0, 432.1, 432.2, 432.9, 851-854</td>
</tr>
</tbody>
</table>

Key: ICD-9-CM – International Classification of Diseases, Ninth Revision, Clinical Modification

Table 3. Bleeding category of bleeding events in patients on warfarin, and warfarin plus ginkgo combination, after limiting follow-up to one year and dropping bleeding events on first day of followup (consistent with the Figure 2 graph)

<table>
<thead>
<tr>
<th>bleeding category (see Appendix 1)</th>
<th>n (% of 122,827 bleeding events)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>83,802 (68.2)</td>
</tr>
<tr>
<td>8</td>
<td>26,066 (21.2)</td>
</tr>
<tr>
<td>16</td>
<td>5,793 (4.7)</td>
</tr>
<tr>
<td>12</td>
<td>3,759 (3.1)</td>
</tr>
<tr>
<td>11</td>
<td>2,682 (2.2)</td>
</tr>
<tr>
<td>7</td>
<td>2,396 (2.0)</td>
</tr>
<tr>
<td>14</td>
<td>1,764 (1.4)</td>
</tr>
<tr>
<td>10</td>
<td>808 (0.7)</td>
</tr>
<tr>
<td>6</td>
<td>541 (0.4)</td>
</tr>
<tr>
<td>15</td>
<td>372 (0.3)</td>
</tr>
<tr>
<td>13</td>
<td>364 (0.3)</td>
</tr>
<tr>
<td>1</td>
<td>252 (0.2)</td>
</tr>
<tr>
<td>2-5, 17</td>
<td>0</td>
</tr>
</tbody>
</table>

Statistical Analysis of Ginkgo-Warfarin Interaction

The comparison of bleeding events between the warfarin only and the warfarin plus Ginkgo groups was made using Cox regression. The cumulative bleeding risk is displayed graphically with Kaplan-Meier plots. Multivariable Cox regression models were also fitted, controlling in a fixed covariate fashion for several binary comorbidities: age 75 or older, history (Hx) of heart failure, Hx of high blood pressure, Hx of vascular disease, Hx of stroke, Hx of diabetes, Hx of hypertension, Hx of renal disease, Hx of liver disease, and Hx of alcohol use. These comorbidities were identified using ICD codes.

Results

At least one bleeding event was noted in 143,360 of the n=796,396 non-ginkgo, warfarin only, patients (18.0%) and in 2,484 of the n=11,003 warfarin patients who at some point while on warfarin were also noted to be using ginkgo (22.6%). It was discovered, however, that the first bleeding event was most frequently noted on the first day that
warfarin was noted in the EMR (24.4% of non-ginkgo patient with bleeding event, and 16.4% of ginkgo patients with bleeding event). The bleeding events after day 1 were relatively uniform across the follow-up period. (Table 4)

It is not likely the bleeding would occur so frequently on the first day of warfarin use, since warfarin does not have that rapid of an anticoagulant effect. Apparently, many patients that were already on warfarin went to a Veterans Administration (VA) hospital to seek care for a bleeding event. These could have been patients who already had an EMR encounter at the hospital, or a patient who sought care there for the first time. Either way, the bleeding event would give the providers cause to inquire about warfarin use and to note its use in the EMR. So on that day, many of these patients had warfarin use noted in their EMR for the first time. To reduce this measurement bias, the first 30 days of warfarin use (first 30 days of follow-up) were next eliminated from the dataset, which reduced the sample size somewhat. This also insured that warfarin dose was stable at a level acceptable to the provider and warfarin induced anticoagulation had reached a therapeutic level. This effectively eliminated the uncharacteristic spike in bleeding events at the first day of follow-up, which was now day 31 on warfarin. (Table 4)

Table 4. Timing of first bleeding event and first mention of ginkgo in EMR

<table>
<thead>
<tr>
<th></th>
<th>Warfarin Only (non-ginkgo group)</th>
<th>Warfarin + Ginkgo (ginkgo group)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (%) of group size</td>
<td>143,360 (18.0)</td>
<td>2,484 (22.6)</td>
</tr>
<tr>
<td>Three most frequent days-on-warfarin when first bleeding event occurred, n (% of bleeding events)</td>
<td>day 1: 34,951 (24.4)</td>
<td>day 1: 408 (16.4)</td>
</tr>
<tr>
<td></td>
<td>day 2: 301 (0.2)</td>
<td>day 4: 7 (0.3)</td>
</tr>
<tr>
<td></td>
<td>day 8: 219 (0.2)</td>
<td>day 21: 7 (0.3)</td>
</tr>
<tr>
<td>Three most frequent days-on-warfarin when ginkgo use first noted, n (% of group size)</td>
<td>day 1: 3,033 (27.6)</td>
<td>day 2: 41 (0.4)</td>
</tr>
<tr>
<td></td>
<td>day 20: 35 (0.3)</td>
<td></td>
</tr>
<tr>
<td>After eliminating first 30 days of warfarin use</td>
<td>Warfarin Only (non-ginkgo group)</td>
<td>Warfarin + Ginkgo (ginkgo group)</td>
</tr>
<tr>
<td>n (%) of group size</td>
<td>122,964 (17.2)</td>
<td>392 (4.1)</td>
</tr>
<tr>
<td>Three most frequent days-on-warfarin when first bleeding event occurred</td>
<td>day 36: 418 (0.3)</td>
<td>day 33: 3 (0.8)</td>
</tr>
<tr>
<td></td>
<td>day 31: 408 (0.3)</td>
<td>day 34: 7 (0.8)</td>
</tr>
<tr>
<td></td>
<td>day 35: 394 (0.3)</td>
<td>day 42: 6 (0.8)</td>
</tr>
<tr>
<td>Three most frequent days-on-warfarin when ginkgo use first noted, n (% of group size)</td>
<td>day 36: 47 (0.5)</td>
<td>day 31: 45 (0.5)</td>
</tr>
<tr>
<td></td>
<td>day 51: 41 (0.4)</td>
<td></td>
</tr>
</tbody>
</table>

Next, we created a new day 1 of follow-up. For the non-ginkgo group, this was day 31, recoded to day 1. For the ginkgo group, it was the first day that ginkgo was noted in the EMR after the first 30 days of warfarin use. For some ginkgo patients, this was warfarin day 31, and for others, it was somewhat uniform across days after that. The first day was recoded to day 1 in this group, as well. The follow-up, then, represented the days on warfarin only for the non-ginkgo group, and days on warfarin + ginkgo combination for the ginkgo group. Follow-up ended with the first bleeding event, or the end of therapy of these agents in the EMR, whichever came first. This is consistent with a time to first event survival analysis (Cox regression).

To investigate if a bleeding event created an information bias, where the provider would more carefully inquire about ginkgo use if a bleeding event occurred, we created a table in a case-control fashion (Table 5). If bleeding
created an obvious information bias, giving the provider cause to inquire about ginkgo use, ginkgo would be expected to be noted in the EMR more often than in patients who never have a bleeding event. The opposite occurred. Ginkgo was noted one-fifth as often in patients with at least one bleeding event compared to patients without a bleeding event (Table 5). This reveals that a ginkgo information bias is only subtle, if it does exist.

Table 5. Probability (%) of noting Ginkgo in the EMR, conditional upon a bleeding event was also noted in the EMR.

<table>
<thead>
<tr>
<th>At least one bleeding event noted in the EMR</th>
<th>Ginkgo noted at least once in the EMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>9209/602916 (1.5%)</td>
</tr>
<tr>
<td>Yes</td>
<td>392/123356 (0.3%)</td>
</tr>
</tbody>
</table>

The follow-up was limited to one year, under the assumption that if ginkgo interacts with warfarin, its effect should be seen by then. It was suspected that ginkgo might first appear in the EMR because a bleeding event had occurred, giving the provider reason to more carefully inquire about herbal use with the patient. Table 6 shows the patients with follow-up periods of days 1, 2, or 3, to determine if bleeds are occurring at the time of beginning of follow-up. This was true about one-third of the time in ginkgo group. Unfortunately, there is no way to extract from the EMR when ginkgo exposure actually first occurred, just when its use was first recorded.

Table 6. Days of follow-up after eliminating first 30 days of warfarin use and restricting follow-up to one year

<table>
<thead>
<tr>
<th>Warfarin Only (non-ginkgo group) [ n = 717,831 ]</th>
<th>Warfarin + Ginkgo (ginkgo group) [ n = 9,601 ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>No bleed [ n = 677,855 ]</td>
<td>No bleed [ n = 9,270 ]</td>
</tr>
<tr>
<td>Bleed [ n = 38,816 ]</td>
<td>Bleed [ n = 331 ]</td>
</tr>
<tr>
<td>Follow-up period in days for first three days, n (%)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>880 (0.1)</td>
</tr>
<tr>
<td>2</td>
<td>714 (0.1)</td>
</tr>
<tr>
<td>3</td>
<td>694 (0.1)</td>
</tr>
<tr>
<td></td>
<td>408 (1.0)</td>
</tr>
<tr>
<td></td>
<td>331 (0.9)</td>
</tr>
<tr>
<td></td>
<td>296 (0.8)</td>
</tr>
<tr>
<td></td>
<td>2,721 (29.3)</td>
</tr>
<tr>
<td></td>
<td>70 (0.8)</td>
</tr>
<tr>
<td></td>
<td>34 (0.4)</td>
</tr>
<tr>
<td></td>
<td>121 (36.6)</td>
</tr>
<tr>
<td></td>
<td>5 (1.5)</td>
</tr>
<tr>
<td></td>
<td>2 (0.6)</td>
</tr>
</tbody>
</table>

Ignoring time-at-risk, a bleeding event was observed in in 331 (3.4%) patients in the warfarin plus ginkgo group and in 38,816 (5.4%) patients in the warfarin only group. After accounting for time-at-risk in a univariable Cox regression model, ginkgo was associated with a higher risk of bleeding (hazard ratio = 2.08, 95%CI: 1.87 to 2.32, p<.001). The association is shown as a Kaplan-Meier curve in Figure 1. The association was unchanged after adjusting for co-morbidities (hazard ratio = 2.08, 95%CI: 1.87 to 2.32, p<.001).
To assess if the association was just a matter of the high incidence of bleeding noted on the first day of ginkgo follow-up, the analysis was repeated after dropping subjects with one day of follow-up. This left a sample size of n=6,759 with in the warfarin plus ginkgo group, with 210 bleeding events. Ignoring time-at-risk, a bleeding event was observed in 210 (3.1%) patients in the warfarin plus ginkgo group and in 38,408 (5.4%) patients in the warfarin only group. After accounting for time-at-risk in a univariable Cox regression model, ginkgo was associated with a higher risk of bleeding (hazard ratio = 1.38, 95%CI: 1.20 to 1.58, p<.001). The association was unchanged after adjusting for co-morbidities (hazard ratio = 1.38, 95%CI: 1.20 to 1.58, p<.001).
Figure 2  Kaplan-Meier graph of risk of bleeding for one year of follow-up, after dropping patients with only one day of follow-up, which hazard ratio from a univariable Cox regression model.

**Discussion**

Although the potential bleeding risk of *Ginkgo biloba* has been much discussed in the literature, initial concerns were based on case reports. These reports described a temporal association between ginkgo use and the bleeding events (4). Some evaluations of randomized controlled trials have not found a higher bleeding risk (10). A systematic review and meta analysis of 18 randomized controlled trials looked at the impact of ginkgo on hemostasis parameters associated with bleeding risk and found a significant reduction in blood viscosity. However, there were no effects on other factors, such as ADP-induced platelet aggregation, fibrinogen concentration, activated partial thromboplastin time, and prothrombin time. Thus the authors concluded there was no higher bleeding risk. Another evaluation looked at claims data in the Taiwan National Health Insurance Research Database (11). The authors concluded there was no significant correlation to the risk of hemorrhage but did note that caution should be exercised in elderly persons and those with known bleeding risk.

Thus the existing literature does not confirm the ginkgo-anticoagulant interaction. The effects and true risks of this interaction are difficult to estimate, based on the limited quantity and quality of existing reports. Leveraging a large database, with documentation of herb use in a high number of patients with cardiovascular disease, as is the case with this project, a positive signal between warfarin and ginkgo was found. This is new information that may help to elucidate the relationship between gingko and warfarin usage and bleeding events.

One limitation we noted is that the documentation rate of herb usage in medical records is possibly low. As shown in the Table 7, the NLP extracted ginkgo use in this study is fairly consistent with those reported in the literature.
Table 7. Prevalence of Ginkgo usage derived from the NLP of VHA notes and the prevalence reported by literature.

<table>
<thead>
<tr>
<th>Population</th>
<th>Time Period</th>
<th>Sample Size</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>National sample of adults (12)</td>
<td>Within 12 months</td>
<td>23,393</td>
<td>1.8%</td>
</tr>
<tr>
<td>Nursing home patients (13)</td>
<td>Within 12 months</td>
<td>68,403</td>
<td>0.4%</td>
</tr>
<tr>
<td>Veterans with cancer (14)</td>
<td>Current use</td>
<td>200</td>
<td>0%</td>
</tr>
<tr>
<td>Veteran outpatients (15)</td>
<td>Current use</td>
<td>458</td>
<td>10%</td>
</tr>
<tr>
<td>Veterans with cancer (16)</td>
<td>Current use</td>
<td>200</td>
<td>1%</td>
</tr>
<tr>
<td>Veterans patients on warfarin</td>
<td>Concurrent with warfarin</td>
<td>726,272</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

Another limitation of study is the lack of detailed information in the EMR. The length of time of ginkgo is obviously longer than what is noted in the EMR. This could have impacted the statistical result. However, the separation of the two curves in Figure 2 occurring after day 30 of follow-up helps support the increased risk observed with the warfarin-ginkgo combination. The dosage and frequency of herbal product use is not regulated and rarely recorded. Information bias is well-known to exist in medical records, with missing information and more complete information being in sicker patients (17).

A final limitation is the possibility of confounding by other medications. That is, some patients might have received other medications that could have increased their risk of bleeding, such as aspirin or other antiplatelet agents. No attempt was made in this study to collect those data in order to control for such medications. However, to use confounding by other medications as an explanation for the direction of the observed study outcome, with higher bleeding risk for the warfarin plus ginkgo group, these other medications would have to have been prescribed more often in the warfarin plus ginkgo group, and it is not likely providers would have chosen to do that.

Despite the limitations, our analysis provides information that is previously unavailable to researchers and clinicians. Given the wide use of ginkgo in the US population, the bleeding risks associated with gingko-warfarin usage is worth noting. In the future, we plan to investigate other suspected herb-drug interactions.

**Acknowledgements**

This project is supported by R01LM011334, R01AT006548, HIR 08-374, HIR 08-204, and CRE 12-315, with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 8UL1TR000105 (formerly UL1RR025764).
References

### Appendix 1. Bleeding Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>ICD9 Codes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>530.82</td>
<td>Esophageal hemorrhage</td>
</tr>
<tr>
<td></td>
<td>531.2</td>
<td>Acute gastric ulcer with perforation</td>
</tr>
<tr>
<td>2</td>
<td>531.4</td>
<td>Chronic or unspecified gastric ulcer with hemorrhage</td>
</tr>
<tr>
<td></td>
<td>531.6</td>
<td>Chronic or unspecified gastric ulcer with hemorrhage and perforation</td>
</tr>
<tr>
<td></td>
<td>532.2</td>
<td>Acute duodenal ulcer with hemorrhage and perforation</td>
</tr>
<tr>
<td>3</td>
<td>532.4</td>
<td>Chronic or unspecified duodenal ulcer with hemorrhage</td>
</tr>
<tr>
<td></td>
<td>532.6</td>
<td>Chronic or unspecified duodenal ulcer with hemorrhage and perforation</td>
</tr>
<tr>
<td></td>
<td>533.2</td>
<td>Acute peptic ulcer of unspecified site with hemorrhage and perforation</td>
</tr>
<tr>
<td>4</td>
<td>533.4</td>
<td>Chronic or unspecified peptic ulcer of unspecified site with hemorrhage</td>
</tr>
<tr>
<td></td>
<td>533.6</td>
<td>Chronic or unspecified peptic ulcer of unspecified site with hemorrhage and perforation</td>
</tr>
<tr>
<td></td>
<td>534.2</td>
<td>Acute gastrojejunal ulcer with hemorrhage and perforation</td>
</tr>
<tr>
<td>5</td>
<td>534.4</td>
<td>Chronic or unspecified gastrojejunal ulcer with hemorrhage</td>
</tr>
<tr>
<td></td>
<td>534.6</td>
<td>Chronic or unspecified gastrojejunal ulcer with hemorrhage and perforation</td>
</tr>
<tr>
<td>6</td>
<td>537.83</td>
<td>Angiodysplasia of stomach and duodenum with hemorrhage</td>
</tr>
<tr>
<td></td>
<td>562.02</td>
<td>Diverticulosis of small intestine with hemorrhage</td>
</tr>
<tr>
<td></td>
<td>562.03</td>
<td>Diverticulitis of small intestine with hemorrhage</td>
</tr>
<tr>
<td>7</td>
<td>562.12</td>
<td>Diverticulosis of colon with hemorrhage</td>
</tr>
<tr>
<td></td>
<td>562.13</td>
<td>Diverticulitis of colon with hemorrhage</td>
</tr>
<tr>
<td>8</td>
<td>569.3</td>
<td>Hemorrhage of rectum and anus</td>
</tr>
<tr>
<td>9</td>
<td>578*</td>
<td>Gastrointestinal hemorrhage</td>
</tr>
<tr>
<td>10</td>
<td>430*</td>
<td>Subarachnoid hemorrhage</td>
</tr>
<tr>
<td>11</td>
<td>431*</td>
<td>Intracerebral hemorrhage</td>
</tr>
<tr>
<td></td>
<td>432.0</td>
<td>Nontraumatic extradural hemorrhage</td>
</tr>
<tr>
<td></td>
<td>432.1</td>
<td>Subdural hemorrhage</td>
</tr>
<tr>
<td></td>
<td>432.2</td>
<td>&lt;non-existent&gt;</td>
</tr>
<tr>
<td></td>
<td>432.9</td>
<td>Unspecified intracranial hemorrhage</td>
</tr>
<tr>
<td>12</td>
<td>851*</td>
<td>Cerebral laceration and contusion</td>
</tr>
<tr>
<td>13</td>
<td>852*</td>
<td>Subarachnoid subdural and extradural hemorrhage following injury</td>
</tr>
<tr>
<td>14</td>
<td>853*</td>
<td>Other and unspecified intracranial hemorrhage following injury</td>
</tr>
<tr>
<td>15</td>
<td>854*</td>
<td>Intracranial injury of other and unspecified nature</td>
</tr>
</tbody>
</table>
Recognizing Disjoint Clinical Concepts in Clinical Text Using Machine Learning-based Methods

Buzhou Tang, PhD\textsuperscript{1,2}, Qingcai Chen, PhD\textsuperscript{1}, Xiaolong Wang, PhD\textsuperscript{1}, Yonghui Wu, PhD\textsuperscript{2}, Yaoyun Zhang PhD\textsuperscript{2}, Min Jiang, MS\textsuperscript{2}, Jingqi Wang, MS\textsuperscript{2}, Hua Xu, PhD\textsuperscript{2}

\textsuperscript{1}Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China
\textsuperscript{2}School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

ABSTRACT

Clinical concept recognition (CCR) is a fundamental task in clinical natural language processing (NLP) field. Almost all current machine learning-based CCR systems can only recognize clinical concepts of consecutive words (called consecutive clinical concepts, CCCs), but can do nothing about clinical concepts of disjoint words (called disjoint clinical concepts, DCCs), which widely exist in clinical text. In this paper, we proposed two novel types of representations for disjoint clinical concepts, and applied two state-of-the-art machine learning methods to recognizing consecutive and disjoint concepts. Experiments conducted on the 2013 ShARe/CLEF challenge corpus showed that our best system achieved a “strict” F-measure of 0.803 for CCCs, a “strict” F-measure of 0.477 for DCCs, and a “strict” F-measure of 0.783 for all clinical concepts, significantly higher than the baseline systems by 4.2% and 4.1% respectively.

INTRODUCTION

With rapid growth of medical informatics technology, a large number of electronic health records (EHRs) have been available in recent years, including a huge mass of data, such as clinical narratives. They have been being used not only to support computerized clinical systems (e.g., computerized clinical decision support systems [1][2]), but also to help the development of clinical and translational research [3]. One of the challenges to use them is that much information is embedded in clinical notes, but cannot be directly accessible for computerized clinical systems which rely on structured information. Therefore, natural language processing (NLP) technologies, which can extract structured information from narrative text, have received great attention in medical domain [4], and many clinical NLP systems have been developed for different applications [5].

Clinical concept recognition (CCR) as a fundamental task of clinical NLP has also attracted great attention, and a large number of systems have been developed to recognize clinical concepts from various types of clinical notes in last two decades. Earlier systems were based on rules or dictionaries manually built by medical experts. The representative systems included MedLEE [4], SymText/MPlus [6][7], MetaMap [8], KnowledgeMap [9], cTAKES [10], and HiTEX [11]. In the past few years, with the increasingly available annotated clinical corpora, researchers started to apply machine learning algorithms to CCR, and several clinical NLP challenges were organized to promote the research on this task. For example, the Center for Informatics for Integrating Biology & the Beside (i2b2) organized NLP
challenges on CCR in 2009 and 2010 respectively [12][13]. The ShARe/CLEF eHealth Evaluation Lab (SHEL) organized a NLP challenge in 2013 [14], which includes a subtask of disorder recognition [15], and launched a similar NLP challenge as a part of the SemEval (Semantic Evaluation) in 2014 (i.e., SemEval-2014 Task 7) using the same training set, but different test set. Rule-based, machine learning-based and hybrid methods were developed to recognize medications by over twenty participating teams in the 2009 i2b2 NLP challenge [12]. In the 2010 i2b2 NLP challenge, clinical concepts, including problems, tests and treatments, not limited to medication, were required to recognize [13]. Most systems were primarily based on machine learning algorithms in this challenge, likely due to a large available annotated corpus [14]. In both the 2013 ShARe/CLEF and SemEval-2014 challenges, machine learning-based systems achieved state-of-the-art performance on disorder concept recognition [15]. Among the four NLP challenges, the ShARe/CLEF and SemEval-2014 challenges first considered disjoint clinical concepts (DCCs), which consist of multiple non-consecutive sequences of tokens. Actually, DCCs always drew much attention because they widely exists in clinical text and are important for subsequent applications such as clinical reasoning systems. In the 2013 ShARe/CLEF challenge, CCR is a preliminary step for concept mapping, mapping concepts in clinical text to UMLS concepts. Almost no machine learning-based method was proposed for disjoint clinical concept recognition except some simple rule-based systems such as MetaMap [8] and cTAKES [10], as well as disjoint named entity recognition in other domains. The main reason may lie in that it is more difficult to annotate DCCs than CCCs. In the 2013 ShARe/CLEF challenge, the organizers annotated a corpus with both CCCs and DCCs where DCCs account for about 10%. This corpus gave us a good chance to investigate how to recognize DCCs using machine learning-based methods. Compared with CCC recognition, the main challenge of DCC recognition is how to represent them. In previous studies, clinical concepts were typically represented by “BIO” tags, where B, I and O denotes that a token is at the beginning, inside and outside of a concept respectively [16]. This representation worked very well for CCCs, but not suitable for DCCs. For DCC recognition, a few systems tried some methods in the 2013 ShARe/CLEF challenge, including rule-based [17] and machine learning-based [18][19] [20] methods. The machine learning-based methods showed much better performance than the rule-based methods. Among all machine learning-based methods, our method was the best [14], which proposed a novel representation for DCCs.

In this paper, we proposed another type of representation for DCCs based on our previous work for the 2013 ShARe/CLEF challenge, which ranked first among 20 participating teams [14][15] on the disorder recognition task. In that study, we proposed a novel type of representation for DCCs by using two additional tags (i.e., H-head entity and D-non-head entity) based on the representation for CCCs, and a two-step method to recognize them. Although this type of representation can completely separate DCCs from CCCs, it does not provide enough information about how to combine head entities and non-head entities except that head entities should be shared with more than one DCC, and non-head entities should be combined with other head/non-head entities. Therefore, Some extra rules is needed for head/non-head entity combination. In this study, we proposed another type of representation for DCCs that integrates combination information into the representation strategy. Using this representation, a separate step for combination is not required any more. Similar to our
previous work in [20], we compared Conditional Random Fields (CRF) [21] and Structured Support Vector Machines (SSVM) [22] when using the two types of representations. To prove the effectiveness of our methods, we also compared them with the CRF/SSVM-based systems ignoring DCCs. To the best of our knowledge, it is the first time to comprehensively investigate DCC recognition using machine learning-based methods, which can be used as a benchmark for further studies. Moreover, the methods proposed in this paper can also be easily applicable to recognize disjoint named entities in other domains.

METHODS

Representations for CCCs and DCCs

We adopted “BIO” tags to represent CCCs in the whole paper. For convenience, we did not repeat any more. For DCCs, two different types of representations were proposed. The first representation used “BIOHD”, where ‘H’ denotes head entities which are consecutive sequences of tokens shared by multiple disjoint concepts in a sentence, and ‘D’ denotes non-head entities which are consecutive sequences of tokens in a disjoint concept not shared by other disjoint concepts in a sentence. The second representation used “BIOHD1234”, where ‘1’, ‘2’, ‘3’ and ‘4’ denote that a non-head entity is combined with the nearest head entity at left, the nearest non-head entity at left, the nearest head entity at right and the nearest non-head entity at right respectively. Figure 1 shows some examples of DCCs represented by “BIOHD” and “BIOHD1234”, where three sentences were used to illustrate different cases of DCCs. In sentence 1, two disjoint disorders (i.e., “extremities … turned in” and “extremities … clinched together”) share a head entity (i.e., “extremities”). The head entity is represented by “extremities/HB-disorder” no matter using “BIOHD” or “BIOHD1234”. The non-head entities (i.e., “turned in” and “clinched together”) are represented by “turned/DB-disorder in/DI-disorder” and “clinched/DB-disorder together/DI-disorder” when using “BIOHD”, while are represented by “turned/D1B-disorder in/D1I-disorder” and “clinched/D1B-disorder together/D1I-disorder” when using “BIOHD1234” as both of them should be combined with the nearest head entity at left. Considering that head entities are represented by “H” in both “BIOHD” and “BIOHD1234”, and non-head entities are always represented by “D” when using “BIOHD”, we did not explain them any more in the following examples. All non-head entities mentioned in the following examples are the cases of representing DCCs using “BIOHD1234”. In sentence 2, two disjoint disorders (i.e., “ABD … distend” and “ABD … tenderness … RUQ”) share a head entity (i.e., “ABD”). The non-head entity in the first disjoint disorder (i.e., “distend”) is represented by “D1B-disorder” as it should be combined with the nearest head entity at left. The non-head entities (i.e., “tenderness” and “RUQ”) in the second disjoint disorder are represented by “D1B-disorder” and “D2B-disorder” respectively because “tenderness” should be combined with the nearest head entity at left, while “RUQ” should be combined with the nearest non-head entity at left (i.e., “tenderness”). In sentence 3, there is a disjoint disorder (i.e., “left atrium … dilated”) composed of two non-head entities (i.e., “left atrium” and “dilated”), which is represented by “D4D2…D2” (i.e., “left/D4B-disorder atrium/D4I-disorder” and “dilated/D2B-disorder”) as the first non-head entity (at left) should be combined with the second non-head entity (at right), and vice versa.
Sentence 1: “Her extremities appeared turned in and clinched together except for her right arm.”
Concepts: “extremities... turned in” and “extremities... clinched together”
BIOHD: “Her/O extremities/HB-disorder appeared/O turned/DB-disorder in/DI-disorder and/O clinched/DB-disorder together/DI-disorder except/O for/O her/O right/O arm/O /O”
BIOHDI34: “Her/O extremities/HB Disorder appeared/O turned/D1B-disorder in/D1I-disorder and/O clinched/D1B-disorder together/DI1-disorder except/O for/O her/O right/O arm/O /O”

Sentence 2: “ABD: mild distend, mild tenderness in RUQ.”
Concepts: “ABD... distend” and “ABD... tenderness... RUQ”

Sentence 3: “The left atrium is dilated.”
Concepts: “left atrium... dilated”
BIOHD: “The/O left/D2B-disorder atrium/DI-disorder is/O dilated/DB-disorder /O”

Figure 1. Examples of two different representations for disjoint clinical concepts.

When using ”BIOHD” tags to represent disjoint clinical concepts, it is required to develop another system to determine how to combine head entities and non-head entities in a sentence as mentioned in our previous work [20], where two simple rules are proposed for combination as follows:

1) For each head entity, combine it with all other non-head entities to form disjoint concepts.
2) If no head entity, combine all non-head entities together to form a disjoint concept.

When using “BIOHDI34” tags to represent DCCs, it is straightforward to combine head entities and non-head entities since the combination information has been integrated into tags. For example, in the sentence 2 in Figure 1, there are two paths from non-head entities to the head entities: 1) distend—ABD; 2) RUQ—tenderness—ABD, forming two disjoint disorders.

Machine learning-based consecutive and disjoint CCR

When CCCs and DCCs are represented by “BIOHD” or “BIOHDI34”, CCR can be recognized as a classification problem, where each token of a sentence is labeled with a tag of “BIOHD” or “BIOHDI34” as shown in Figure 1. For “BIOHD”, the tag set is {B-disorder, I-disorder, O, HB-disorder, HI-disorder, DB-disorder, DI-disorder}. For “BIOHDI34”, the tag set is {B-disorder, I-disorder, O, HB-disorder, D1B-disorder, D1I-disorder, D2B-disorder, D2I-disorder, D3B-disorder, D3I-disorder, D4B-disorder, D4I-disorder}. Moreover, the machine learning methods proposed for consecutive CCR can also be applied to consecutive and disjoint CCR, such as Support Vector Machine (SVM) [13], CRF [23], SSVM [24][25] and so on. In our study, we compared CRF and SSVM, two state-of-the-art machine learning methods for consecutive and disjoint CCR [25], and used CRFsuite (http://www.chokkan.org/software/crfsuite/) and SVMhmm (http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html) as implementations of CRF and SSVM respectively. For sentence boundary detection and tokenization, we adopted the corresponding modules of MedEx (https://code.google.com/p/medex-uima/downloads/list), a specific tool for medical information extraction.
**Features**

We used the same features as our previous work in [20], including bag-of-word, part-of-speech (POS), note type, section information, word representations, semantic categories of words. Most of them were also used in our previous systems for medical concept recognition [23][25][24][26]. For detailed information, please refer to the references.

**Dataset**

We used the dataset of the 2013 ShARe/CLEF challenge, which is composed of 298 notes from different clinical encounters including radiology reports, discharge summaries, ECG reports and ECHO reports. For each note, only disorders, including consecutive and disjoint disorders, were annotated according to a pre-defined guideline. The dataset was divided into two parts: a training set of 199 notes used for system development, and a test set of 99 notes used for system evaluation. In the training set, 651 out of 5811 disorders were disjoint, and 439 out of 5340 disorders were disjoint in the test set. Table 1 shows the counts of disorders in the training and test sets, where the number in parenthesis is the count of disjoint disorders.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>#Note</th>
<th>#Concept</th>
<th>#Consecutive</th>
<th>#Disjoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>All</td>
<td>199</td>
<td>5816</td>
<td>5165</td>
<td>651</td>
</tr>
<tr>
<td></td>
<td>ECHO</td>
<td>42</td>
<td>828</td>
<td>603</td>
<td>225</td>
</tr>
<tr>
<td>Training</td>
<td>RADIOLOGY</td>
<td>42</td>
<td>555</td>
<td>489</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>DISCHARGE</td>
<td>61</td>
<td>3589</td>
<td>3232</td>
<td>357</td>
</tr>
<tr>
<td></td>
<td>ECG</td>
<td>54</td>
<td>193</td>
<td>190</td>
<td>3</td>
</tr>
<tr>
<td>All</td>
<td>All</td>
<td>99</td>
<td>5340</td>
<td>4901</td>
<td>439</td>
</tr>
<tr>
<td></td>
<td>ECHO</td>
<td>12</td>
<td>338</td>
<td>280</td>
<td>58</td>
</tr>
<tr>
<td>Test</td>
<td>RADIOLOGY</td>
<td>12</td>
<td>162</td>
<td>158</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>DISCHARGE</td>
<td>75</td>
<td>4840</td>
<td>4463</td>
<td>377</td>
</tr>
<tr>
<td></td>
<td>ECG</td>
<td>0</td>
<td>0 (0)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Experiments and Evaluation**

All models were trained on the training set and evaluated on the test set, and their parameters were optimized by 10-fold cross-validation on the training set. The performance of disorder concept recognition were evaluated by precision (P), recall (R) and F-measure (F) in both “strict” and “relaxed” modes [14], calculated by an evaluation tool provided by the organizers ([https://sites.google.com/site/shareclefehealth/evaluation](https://sites.google.com/site/shareclefehealth/evaluation)). To investigate the effect of the proposed methods, we started with two types of baseline systems: one treated every head and non-head entity of disjoint clinical concepts as an individual CCC (1st); and the other one removed all DCCs (2nd), and then compared them with our systems.
RESULTS

Table 2 shows the overall performance of the machine learning-based consecutive and disjoint CCR systems on the test set, when using different types of representations for CCCs and DCCs. The systems using “BIOHD” or “BIOHD1234” showed significantly better performance than the corresponding baseline systems, indicating that the proposed representations are suitable for CCCs and DCCs. For example, when using “BIOHD”, the CRF-based system outperformed the 1st baseline system by 3.9% (0.777 vs 0.738), and the 2nd baseline system by 5.9% (0.777 vs 0.718) in “strict” F-measure respectively; The SSVM-based system outperformed the 1st and 2nd baseline systems by 4.1% (0.782 vs 0.741) and 4.0% (0.782 vs 0.742) in “strict” F-measure respectively. When using the same representations for CCCs and DCCs, the SSVM-based systems showed better performance than the CRF-based systems. For example, the “strict” F-measure of the SSVM-based system was 0.782, while that of the CRF-based system was 0.777 when using “BIOHD”. For each machine learning method, the system using “BIOHD1234” slightly outperformed the system using “BIOHD”. For example, the SSVM-based system using “BIOHD1234” achieved a “strict” F-measure of 0.783, while the SSVM-based system using “BIOHD” achieved a “strict” F-measure of 0.782.

Here, we should note that the CRF-based and SSVM-based systems using “BIOHD” in this paper achieved better performance than our previous systems submitted to the 2013 ShARe/CLEF challenge as some bugs in previous systems have been fixed in current ones.

Table 2. Overall performance of the machine learning-based CCR systems.

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Baseline</td>
<td>CRF</td>
<td>0.773</td>
<td>0.707</td>
<td>0.738</td>
<td>0.937</td>
<td>0.848</td>
</tr>
<tr>
<td></td>
<td>SSVM</td>
<td>0.764</td>
<td>0.720</td>
<td>0.741</td>
<td>0.933</td>
<td>0.863</td>
</tr>
<tr>
<td>2nd Baseline</td>
<td>CRF</td>
<td>0.862</td>
<td>0.615</td>
<td>0.718</td>
<td>0.965</td>
<td>0.693</td>
</tr>
<tr>
<td></td>
<td>SSVM</td>
<td>0.842</td>
<td>0.663</td>
<td>0.742</td>
<td>0.947</td>
<td>0.749</td>
</tr>
<tr>
<td>BIOHD</td>
<td>CRF</td>
<td>0.839</td>
<td>0.723</td>
<td>0.777</td>
<td>0.952</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td>SSVM</td>
<td>0.830</td>
<td>0.740</td>
<td>0.782</td>
<td>0.941</td>
<td>0.849</td>
</tr>
<tr>
<td>BIOHD1234</td>
<td>CRF</td>
<td>0.845</td>
<td>0.722</td>
<td>0.778</td>
<td>0.955</td>
<td>0.831</td>
</tr>
<tr>
<td></td>
<td>SSVM</td>
<td>0.834</td>
<td>0.739</td>
<td>0.783</td>
<td>0.942</td>
<td>0.846</td>
</tr>
</tbody>
</table>

Furthermore, table 3 shows the performance of machine learning-based systems in the “strict” mode for CCCs and DCCs respectively. The machine learning-based system using “BIOHD” or “BIOHD1234” not only correctly recognized a number of DCCs, but also improved the performance for CCCs. For example, the “strict” F-measure of the CRF-based system using “BIOHD” for DCCs was 0.433, and that for CCCs was 0.799, higher than the 1st and 2nd baseline systems by 2.8% (0.799 vs 0.771) and 4.4% (0.799 vs 0.753). When using the same
representations for CCCs and DCCs, the SSVM-based systems showed better performance than CRF-based systems for both CCCs and DCCs. For example, when using “BIOHD”, the SSVM-based system achieved a “strict” F-measure of 0.802 for CCCs and a “strict” F-measure of 0.487 for DCCs; while the CRF-based system achieved a “strict” F-measures of 0.799 and 0.433 for CCCs and DCCs respectively. For each machine learning method, the systems using “BIOHD1234” showed slightly better performance than the systems using “BIOHD” for CCCs, but slightly worse performance for DCCs. For each machine learning-based system, the performance for CCCs is much higher than that for DCCs.

Table 3. Performance of the machine learning-based systems for CCCs and DCCs (“strict”).

<table>
<thead>
<tr>
<th>System</th>
<th>CCCs</th>
<th>DCCs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>1st Baseline</td>
<td>CRF</td>
<td>0.773</td>
</tr>
<tr>
<td></td>
<td>SSVM</td>
<td>0.764</td>
</tr>
<tr>
<td>2nd Baseline</td>
<td>CRF</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td>SSVM</td>
<td>0.842</td>
</tr>
<tr>
<td>BIOHD</td>
<td>CRF</td>
<td>0.844</td>
</tr>
<tr>
<td></td>
<td>SSVM</td>
<td>0.832</td>
</tr>
<tr>
<td>BIOHD1234</td>
<td>CRF</td>
<td>0.849</td>
</tr>
<tr>
<td></td>
<td>SSVM</td>
<td>0.834</td>
</tr>
</tbody>
</table>

**DISCUSSION**

It is reasonable that the proposed two types of representations improved the performance of CCR systems as expected since they can distinguish DCCs from CCCs. To evaluate their representation abilities, we calculated upper boundaries of systems using “BIO”, “BIOHD” and “BIOHD1234” on the training set respectively. Firstly, we represented notes with gold clinical concepts using one of the three types of representations, and then converted the labeled notes back to clinical concepts. Finally, the upper boundary was calculated by comparing the converted clinical concepts with the gold ones using the evaluation tool. The upper boundaries of the systems using the three types of representations are shown in Table 3. The upper boundaries of systems using our representations were much higher than the systems using “BIO” (i.e., baseline systems shown in figure 1), and the differences ranged from 6.7% to 12.9% in “strict” F-measure. Among the proposed two types of representations, the upper boundary of the systems using “BIOHD1234” is slightly higher than the systems using “BIOHD” (0.969 vs 0.965 in “strict” F-measure). It means that the representation ability of “BIOHD1234” is slightly stronger than “BIOHD”, and is much stronger than “BIO”.

Although both “BIOHD” and “BIOHD1234” have good ability to represent both CCCs and DCCs, they are not complete. When clinical concepts are represented by “BIOHD”, clinical
Table 3. Upper boundaries of the systems using the three representations on the training set.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Strict P</th>
<th>Strict R</th>
<th>Strict F</th>
<th>Relaxed P</th>
<th>Relaxed R</th>
<th>Relaxed F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st “BIO”</td>
<td>0.806</td>
<td>0.877</td>
<td>0.840</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>2nd “BIO”</td>
<td>0.999</td>
<td>0.816</td>
<td>0.898</td>
<td>1.000</td>
<td>0.817</td>
<td>0.899</td>
</tr>
<tr>
<td>“BIOHD”</td>
<td>0.967</td>
<td>0.962</td>
<td>0.965</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>“BIOHD1234”</td>
<td>0.976</td>
<td>0.963</td>
<td>0.969</td>
<td>1.000</td>
<td>0.997</td>
<td>0.999</td>
</tr>
</tbody>
</table>

concepts in the following three cases might be wrongly represented: 1) more than two head entities in a sentence; 2) more than one non-head entity in DCCs; 3) more than one DCC in a sentence. Some examples of clinical concepts wrongly represented by “BIOHD” are shown in figure 2. When clinical concepts are represented by “BIOHD1234”, the clinical concepts in sentence 2 and 3 are correctly represented, but the clinical concepts in sentence 1 are wrongly represented. Therefore, designing a complete representation for CCCs and DCCs is still a valuable topic, which is one case of our future work.

Sentence 1: “There is a small amount of blood seen within the third and fourth ventricles.”
Gold: “blood … thirdventricles” and “blood … four … ventricles”
BIOHD: “There/O is/O a/O small/C amount/O of/O blood/HB-disorder seen/O within/O the/O third/D3-disorder and/O fourth/D4-disorder ventricles/HB-disorder./O”
Converted: “blood … third”, “blood … fourth”, “third … ventricles” and “fourth … ventricles”

Sentence 2: “**ABD**: mild distend, mild tenderness in RUQ.”
Gold: “**ABD** … distend” and “**ABD** … tenderness … RUQ”
BIOHD: “**ABD/HB-disorder** /O mild/O distend/D4-disorder,/O mild/O tenderness/D4-disorder in/O RUQ/D4-disorder ./O”
Converted: “**ABD** … distend”,”**ABD** … tenderness” and “**ABD** … RUQ”

Sentence 3: “The patient had blood in his mouth and on his tongue, pupils were pinpoint and reactive.”
Gold: “blood … on his tongue” and “pupils … reactive”
BIOHD: “The/O patient/O had/O blood/D4-disorder in/O his/O mouth/O and/O on/D4-disorder : his/DI-disorder tongue/DI-disorder,/O pupils/D4-disorder were/O pinpoint/O and/O reactive/D4-disorder ./O”
Converted: “blood … on his tongue … pupils … reactive”

**Figure 2.** Examples of clinical concepts wrongly represented by “BIOHD”.

Since the ratio of CCCs to DCCs in the 2103 ShARe/CLEF challenge corpus is about 9:1, which is imbalanced, it is easy to understand that the performance of machine learning-based systems for DCCs is not as good as the performance for CCCs. An interesting direction is to consider data imbalance for further improvement, which is another case of our future work.

Moreover, all methods proposed in this paper are not limited to clinical concepts, but also suitable for general named entities. Therefore, they are potentially useful in other domains.

**CONCLUSIONS**

In this study, we investigated the machine learning-based methods to recognizing CCCs and DCCs. Two novel types of representations were proposed for CCCs and DCCs, and their
effectiveness was proved on a benchmark dataset.

Acknowledgements

This study is supported in part by grants: National 863 Program of China (2015AA015405), NSFCs (National Natural Science Foundation of China) (61402128, 61473101, 61173075 and 61272383) and Strategic Emerging Industry Development Special Funds of Shenzhen (ZDSY20120613125401420, JCYJ20140508161040764, JCYJ20140417172417105 and JCYJ20140627163809422).

References


Challenges and Solutions in Optimizing Execution Performance of a Clinical Decision Support-Based Quality Measurement (CDS-QM) Framework

Tyler J Tippetts, Phillip B Warner, MS, Polina V Kukhareva, MPH, David E Shields, Catherine J Staes, MPH, PhD, Kensaku Kawamoto, MD, PhD, MHS
Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

Given the close relationship between clinical decision support (CDS) and quality measurement (QM), it has been proposed that a standards-based CDS Web service could be leveraged to enable QM. Benefits of such a CDS-QM framework include semantic consistency and implementation efficiency. However, earlier research has identified execution performance as a critical barrier when CDS-QM is applied to large populations. Here, we describe challenges encountered and solutions devised to optimize CDS-QM execution performance. Through these optimizations, the CDS-QM execution time was optimized approximately three orders of magnitude, such that approximately 370,000 patient records can now be evaluated for 22 quality measure groups in less than 5 hours (approximately 2 milliseconds per measure group per patient). Several key optimization methods were identified, with the most impact achieved through population-based retrieval of relevant data, multi-step data staging, and parallel processing. These optimizations have enabled CDS-QM to be operationally deployed at an enterprise level.

Introduction

Close relationship between clinical decision support (CDS) and quality measurement (QM). CDS and QM are closely related, in that CDS may recommend that X be done if Y is true, while QM may measure if X was done when Y was true in order to assess the quality of the care given. For example, a CDS rule may recommend that aspirin be given if a patient has suffered an acute myocardial infarction, while a QM may measure if aspirin was given when a patient had suffered an acute myocardial infarction. Another commonality between CDS and QM is the goal of improving the quality of care. It has also been reported in the literature that they also share many of the same implementation challenges.

Given the close relationship between CDS and QM, there have been increasing efforts to align these two related aspects of quality improvement. Of particular interest from an informatics perspective, the U.S. Office of the National Coordinator for Health IT (ONC) and the Centers for Medicare & Medicaid Services (CMS) are sponsoring a public-private partnership known as the Clinical Quality Framework initiative to develop a common set of harmonized interoperability standards to enable both CDS and QM.

Benefits of a CDS-based framework for QM and prototype implementation. Due to the similar nature of CDS and QM, there are at least two important benefits to using a CDS-based quality measurement (CDS-QM) framework, in which the same approach is leveraged for both CDS and QM purposes. The first important benefit is semantic consistency. Because the same technical framework and core clinical knowledge is re-used across both CDS and QM, clinicians can receive consistent assessments both when receiving CDS at the point of care and obtaining performance feedback through QM. The second important benefit to CDS-based QM is implementation efficiency. By re-using the same underlying infrastructure and content, what may otherwise have involved two separate teams and development efforts can now be completed by a single team and effort. Such de-duplication of effort is particularly important given the significant complexity and attendant resources involved in the development, validation, and maintenance of this type of knowledge resource.

To harness these benefits of CDS-based QM, we have previously implemented a prototype CDS-QM framework in which a standards-based CDS Web service is leveraged across a population to enable QM. This earlier prototype was able to successfully automate the evaluation of an inpatient quality measure (SCIP-VTE-2) with a recall of 100% and a precision of 96%.

Need for optimizing execution performance. Despite the benefits of CDS-based QM, earlier research has identified execution performance as a critical challenge to operationally deploying such an approach at scale. For example,
our earlier prototype of this approach\(^6\) required approximately 5 seconds to process a measure group for a single patient. Similarly, the CDS Consortium has previously reported that a similar CDS Web service approach to patient evaluation required a similar amount of time to process.\(^7\) While reasonable for QMs with limited population sizes, such as an inpatient quality measure which may only apply to several hundred patients at any given time, this approach is not scalable for QMs that apply to much larger patient populations, such as outpatient quality measures related to preventive care. For example, at 5 seconds per patient per measure group, processing 22 outpatient quality measure groups for 370,000 patients (e.g., to assess all patients seen within the past year at our health care system) would take over a year to process. Execution time is also an important consideration for real-time CDS as well as population-based CDS involving large numbers of patients.

At University of Utah Health Care (UUHC), our objective was to implement a CDS-QM framework for enterprise deployment, including for outpatient quality measures. To enable such enterprise use of this promising approach to QM, we needed to be able to evaluate a large number of QMs for very large patient populations in less than 24 hours (i.e. within one business day). This manuscript describes the performance optimizations that were implemented in order to meet this business need, such that the processing time was decreased approximately three orders of magnitude, from approximately 5 seconds per measure group per patient to 2 milliseconds per measure group per patient. By describing these performance optimizations and their relative impact, we seek to facilitate others’ operational use of a CDS-QM approach for large-scale quality measurement and population health management.

**Methods**

**Study Context**

This study was conducted to meet operational quality measurement needs at University of Utah Health Care (UUHC), which serves as the Intermountain West’s only academic healthcare system and includes four hospitals, 10 community clinics, over 10,000 employees, and over 1,200 physicians. Approximately 370,000 patients are seen at UUHC in any given year. This study was approved by the University of Utah IRB (protocol # 00080838).

The CDS-QM framework was initially applied at UUHC for calculating inpatient quality measures such as the Patient Safety Indicators (PSI) from the Agency for Healthcare Research & Quality (AHRQ) and Surgical Care Improvement Project (SCIP) measures from the Joint Commission. Because only several hundred hospitalized patients would be relevant for these measures at any given time, execution performance did not need to be optimized for evaluating these measures. The initial CDS-QM framework developed for inpatient quality measurement purposes represents the Phase I framework described below.

Subsequently, there was an institutional need to evaluate Healthcare Effectiveness Data and Information Set (HEDIS) outpatient quality measures from the National Committee for Quality Assurance. HEDIS measures are used by more than 90 percent of U.S. health insurers as a measure of quality and encompass measures such as the frequency of prenatal care visits compared to established standards and the degree of glycemic control in patients with diabetes mellitus.\(^8\) HEDIS measures are calculated based on a combination of administrative/billing data and EHR data such as laboratory results, vital signs, immunization records, and procedure records.

At UUHC, the HEDIS measures had been calculated manually, in part due to the highly complex nature of many of the measures. For example, to calculate the number of prenatal visits that occurred in the first trimester of pregnancy, HEDIS defines 17 distinct patterns of eligibility. For this initiative, 22 measure groups (e.g., comprehensive diabetes care) encompassing 78 individual measures (e.g., % of members 18-75 years of age who had proper Hemoglobin A1c testing) were prioritized for automation.

The initial focus for the CDS-QM effort was to evaluate HEDIS measures for approximately 20,000 patients whom our commercial health plans considered to be primarily managed by UUHC clinicians. The platform developed to handle this scope of evaluation is reflected in the Phase II CDS-QM framework below.

Subsequently, the University of Utah Medical Group decided to base clinician compensation in part on their performance on HEDIS quality measures. This increased the applicable patient population to the approximately 370,000 patients seen at UUHC at least once within a one year timeframe. The platform developed to handle this
scope of evaluation represents the Phase III CDS-QM framework below. In all cases, the business need was for the evaluation of all HEDIS quality measures to be completed for all relevant patients within one business day.

**Required Computational Tasks**

The CDS-QM framework was required to perform three primary computational tasks, each of which served as targets for performance optimization. The first task was data collection, in which patient data were retrieved from the data warehouse. In all phases, query parameters were set to restrict the data to those relevant for the required quality measure evaluations. The second task was data transformation, in which the collected data were converted into the Health Level 7 (HL7) Virtual Medical Record (vMR) data format, which is a standard data model for CDS. Finally, the third task was data evaluation and storage, in which the OpenCDS decision support Web service generates quality measurement assessments based on the vMR data, which are in turn stored in the data warehouse for analysis and reporting.

**Qualitative Description of CDS-QM Framework by Phase**

As described above, the business need for the CDS-QM framework evolved over three phases: a framework needed for evaluations of less than 1,000 patients at a time, evaluations of approximately 20,000 patients at a time, and evaluations of approximately 370,000 patients at a time. Accordingly, the framework was adapted in a phased approach. Each of these phases is described with respect to its approach to the three computational tasks.

**Identification of Challenges and Solutions**

The challenges encountered in optimizing execution performance are described, along with optimization solutions devised. Many different optimization approaches were implemented, but only those with significant impact are discussed here.

**Scalability Analysis for Phase III Framework**

For the final, phase III CDS-QM framework, we measured the time required for evaluating 22 HEDIS quality measure groups for increasingly large patient populations. The 22 HEDIS quality measure groups included 78 individual quality measurements. These analyses were conducted for 1 patient, 1,000 patients, 20,000 patients, and 370,000 patients.

**Results**

**Qualitative Description of CDS-QM Framework by Phase**

The overall CDS-QM process during phases I, II, and III are summarized in Figures 1, 2, and 3 respectively and described below. Moreover, the Table summarizes the optimization challenges, corresponding solutions, and impact for phases II and III. These challenges, solutions, and impact are also discussed in the framework descriptions below.

**Phase I CDS-QM Framework**

*Overall Approach and Performance*

In this phase, all three computational tasks were conducted on a patient-by-patient basis in sequence. The CDS-QM framework averaged approximately 5 seconds per patient per measure group, with almost all of the time required for data collection.
In this phase, data were collected patient-by-patient, encounter-by-encounter using an integration engine called Mirth Connect. The integration engine directly queried the data warehouse and provided a scripting-based platform for accessing and transforming the patient data. An initial query identified the patient cohort to be analyzed. Then, for each identified patient, all subsequent tasks were conducted in sequence on a patient-by-patient basis. With regard to patient data collection, these patient-specific tasks included retrieval of demographic data and retrieval of relevant encounters. Then, for each relevant encounter, relevant associated data, such as for procedures, lab results, diagnoses, medications, and vital signs were retrieved in a sequential, repeating manner. This strategy produced a nested looping process where every nested query was run once for each patient and for each encounter. The collected data were held in volatile memory.

After all queries were completed for an individual patient, the collected data were transformed into HL7 vMR objects. Data transformation occurred sequentially, one patient at a time.

The data were sent using an HL7 Decision Support Service-compliant interface, specifying which quality measures should be used to process the patient vMR. The results for each specified quality measure were returned to and parsed by the integration engine and written back to the data warehouse. This was also done sequentially, one patient after the other.

**Phase II CDS-QM Framework**

Challenges from Previous Phase, Overview of Solutions Devised, and Performance

There were two main challenges from the Phase I framework. First, data retrieval was a clear rate limiting step, such that the approach clearly could not scale to even tens of thousands of patients. A second challenge was that volatile memory limitations of the computing hardware prevented the simultaneous processing of a large number of patients.

To address these challenges, we altered the database queries to obtain data for the entire cohort of interest once, instead of for each patient sequentially. We also parallelized data queries and stored vMR fragments (e.g., vMR
procedure fragments, vMR medication fragments) on disk to allow this parallelization within volatile memory constraints. This phase of the CDS-QM framework allowed approximately 20,000 patients to be evaluated for 22 HEDIS quality measure groups in approximately 8 hours, or an average of about 70 milliseconds per patient per measure group.

**Phase II, Task 1: Patient Data Collection**

The step for collecting patient data was first optimized by modifying and optimizing the query strategy to retrieve data for all patients in the measurement cohort rather than for a single patient per query. As long as source tables were properly indexed, most queries for the population data were nearly as fast as the same queries for a single patient. As a result, significant performance gains were realized after implementing this population-based query approach.

The result sets for these queries were large, so they could not be stored in volatile memory and hence were cached into a temporary schema. The cached results were later processed using a result set cursor, so that only a small portion of the result set was brought into volatile memory at any given time. We also found that our queries could run independently of each other, and hence modified our query strategy to include running the queries in parallel.

**Phase II, Task 2: Data Transformation**

The data transformation step was integrated into the data collection process. Fragments of the vMR were created for each relevant data element type (e.g., procedures, medications, lab results), and these fragments were stored in the temporary data cache as character large objects (CLOBs)\(^2\). The vMR fragments were then compiled with the demographic and encounter data from the temporary cache and transformed into individual patient vMRs.

**Phase II, Task 3: Data Evaluation and Storage**

Directly following vMR creation, each vMR was sent to the CDS Web service for processing. This step remained unmodified from the Phase 1 implementation.

**Phase III CDS-QM Framework**

**Challenges from Previous Phase, Overview of Solutions Devised, and Performance**

There were two main challenges from the Phase II framework. First, storing the transformed vMR fragments on disk became rate-limiting, due to the relatively slow speed of writing CLOBs into the data warehouse and reading

---

1198
from it. Also, for very large runs (e.g., 370,000 patients), this approach could not scale to the amount of storage space available in the data warehouse for this project. A second challenge was the sequential nature of vMR creation and evaluation. Whereas data retrieval was the clear bottleneck in Phase I, data storage (caching) and evaluation became the dominant bottlenecks in Phase II.

To address these challenges, we first stored the raw patient data into a standardized staging schema. For example, instead of storing a vMR XML fragment for procedures as a CLOB, the source data points required to generate this vMR XML fragment were stored in an indexed table with non-CLOB data types such as integers and dates. As a second key solution, we parallelized the vMR creation using multi-threaded processing techniques and used multiple servers, each with a deployed instance of OpenCDS, for evaluation.

Beyond these infrastructure-level optimizations, we also sought to optimize the execution performance of OpenCDS itself. We added the ability to process the set of requested measure groups in parallel. We also improved the performance of multiple requests by scaling the performance by the number of CPU cores (processors) available on the system. This current phase of the CDS-QM framework allows approximately 370,000 patients to be evaluated for 22 HEDIS quality measure groups in less than 5 hours, or an average of about 2 milliseconds per patient per measure group. Of note, continuous validation of CDS-QM results was performed against a sample of manual chart audit results. Systematic improvement based on this continuous validation process has resulted in a system with validity demonstrated to be equivalent to, and in some cases better than, manual QM analysis.

Figure 3. Phase III CDS-QM Process.

Phase III, Task 1: Patient Data Collection
Data collection continued to be conducted using population-based, parallel queries. However, instead of storing the collected data as pre-transformed vMR fragments in CLOBs, the source data elements were stored in their native data formats in a staging database schema.
Phase III, Task 2: Data Transformation

Staging the raw data, as opposed to the vMR fragments, facilitated the decoupling of the transformation task from the data collection task. Another subsystem, hereafter referred to as MeasureRunner, was developed to transform the data from the staging schema to individual patient vMRs. This process allowed the parallel creation of multiple patient vMRs, which in turn enabled the maximal use of available system resources.

Phase III, Task 3: Data Evaluation and Storage

MeasureRunner enabled the parallelization and batch evaluation of vMRs using the OpenCDS decision support service. Multiple instances of the decision support service were deployed to allow parallel processing. Responses were parsed and stored asynchronously as they were returned by the decision support service instances.

Challenges and Solutions

Table 1 summarizes the challenges encountered during this multi-phase optimization process, as well as the key solutions that were found to address these challenges.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Primary Challenges Addressed from Previous Phase</th>
<th>Key Solutions</th>
</tr>
</thead>
</table>
| Phase II | • Data retrieval was clear bottleneck and did not scale to large populations  
• Limitations in the amount of available volatile memory prevented simultaneous processing of large populations | • Altered queries to return data for entire cohort instead of per patient  
• Parallelized data queries  
• vMR fragments stored on disk to allow parallelization |
| Phase III | • Storage of vMR fragments on disk would not scale because of limited disk space and relatively slow speed of reading and writing CLOBs  
• vMR creation and evaluation was sequential | • Stored raw patient data rather than vMR fragments into a standardized staging schema to improve caching and transformation performance  
• Parallelized vMR creation and evaluation using multi-threading and multiple servers |

CLOB = Character Large Object. vMR = Virtual Medical Record.

Scalability Analysis

Figure 4 shows the execution time for the Phase III CDS-QM framework when evaluating increasingly large cohort sizes for 22 HEDIS quality measure groups. As noted, the process scales up well for large cohort sizes, with average completion times per patient per measure group actually decreasing with large cohort sizes.
Discussion

Summary of Findings
A standards-based CDS-QM framework was enhanced across two phases in this study to enable the evaluation of increasingly large patient cohorts in a timely manner. Through key strategies including population-based data collection, parallel processing, and the use of intermediate staging tables, the CDS-QM framework achieved a performance improvement of approximately three orders of magnitude. The current CDS-QM framework is able to process 22 quality measure groups for all patients seen at least once in the past year in an academic health system in less than 5 hours.

Impact
In terms of local impact, the CDS-QM framework now allows complex clinical quality measures to be evaluated for all applicable patients within one business day, as opposed to depending on estimations based on chart audits that required weeks or months to complete. Consequently, the CDS-QM framework is enabling a number of clinical quality improvement initiatives, including those with explicit ties to physician compensation.

In terms of impact beyond our institution, we speculate that the lessons learned from this study will facilitate efforts by others to deliver operational CDS and QM using a common CDS-QM framework, thereby achieving the benefits of semantic consistency and implementation efficiency. In particular, we hope that the open-source nature of the relevant software components will encourage the adoption and community improvement of the CDS-QM framework.

Limitations and Strengths of Study
One significant limitation of this study is that execution timings were not systematically recorded, as the development of each phase of the CDS-QM framework was conducted for operational rather than research purposes. However, we believe that the exact magnitude of benefit gained from each strategy is less important than the overall magnitude of benefit gained from the combination of all strategies. A second limitation of this study is that the execution performance of the CDS-QM framework has not yet been evaluated at other institutions. Finally, the data retrieval component of our implementation is currently institution-specific. While some customization is likely inevitable for institution-specific data retrieval (e.g., to account for differing database schemas), our goal is to generalize this data retrieval component so that institution-specific customizations can be implemented with high efficiency.

An important strength of this study is that the framework described has been shown to be sufficiently robust for operational, enterprise-level use. Indeed, the analysis described in the Phase III framework is now routinely being conducted for operational clinical use. A second strength of this study is that it utilizes open-source, standards-based software components that are freely available. A third strength is the scalability analysis, which indicates that the approach scales well to increasing cohort sizes. A final strength of this approach is that it is a general-purpose framework that can be adopted to other quality measures as well as to other large-scale analysis needs, including population health management.

Future Directions
The core CDS Web service is already freely available through [www.opencds.org](http://www.opencds.org). We intend to make the other components of the CDS-QM framework also freely available after re-factoring the components for general use. We also intend to continually enhance the CDS-QM framework and its knowledge content based on the operational needs of our healthcare system and of other stakeholders making use of this framework.

Conclusion
Leveraging a standards-based CDS infrastructure for quality measurement has important benefits, including semantic consistency and implementation efficiency. We hope that the performance optimization methods identified in this study will enable such a CDS-QM approach to be more widely leveraged in operational settings to improve health and health care.
Acknowledgements

This study was funded by University of Utah Health Care. KK is currently or recently served as a consultant on CDS to the Office of the National Coordinator for Health IT, ARUP Laboratories, McKesson InterQual, ESAC, Inc., JBS International, Inc., Inflexxion, Inc., Intelligent Automation, Inc., Partners HealthCare, the RAND Corporation, and Mayo Clinic. KK receives royalties for a Duke University-owned CDS technology for infectious disease management known as CustomID that he helped develop. KK was formerly a consultant for Religent, Inc. and a co-owner and consultant for Clinica Software, Inc., both of which provide commercial CDS services, including via the use of a CDS technology known as SEBASTIAN that KK developed. KK no longer has a financial relationship with either Religent or Clinica Software. KK has no competing interests related to OpenCDS, which is freely available to the community as an open-source resource. All other authors declare no conflict of interest.
References


Scaling Out and Evaluation of OBSeCAn, an Automated Section Annotator for Semi-Structured Clinical Documents, on a Large VA Clinical Corpus

Le-Thuy T. Tran, PhD, Guy Divita, MS, Andrew Redd, PhD, Marjorie E. Carter, MSPH, Matthew Samore, MD, Adi V. Gundlapalli, MD, PhD, MS
IDEAS Center, VA Salt Lake City Health Care System, Salt Lake City, Utah, USA; Departments of Internal Medicine and Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA

Abstract

“Identifying and labeling” (annotating) sections improves the effectiveness of extracting information stored in the free text of clinical documents. OBSeCAn, an automated ontology-based section annotator, was developed to identify and label sections of semi-structured clinical documents from the Department of Veterans Affairs (VA). In the first step, the algorithm reads and parses the document to obtain and store information regarding sections into a structure that supports the hierarchy of sections. The second stage detects and makes correction to errors in the parsed structure. The third stage produces the section annotation output using the final parsed tree. In this study, we present the OBSeCAn method and its scale to a million document corpus and evaluate its performance in identifying family history sections. We identify high yield sections for this use case from note titles such as primary care and demonstrate a median rate of 99% in correctly identifying a family history section.

Introduction

Information stored in electronic health records (EHRs) such as patient expressed symptoms, physical findings, medications, past medical history, and family history of illnesses has tremendous potential for aiding in the detection of patient care and treatment patterns, risks and outcomes of diseases, or adverse events\(^1,2,3,4\). Effective clinical text mining methods and techniques are important for the discovery of these valuable health data for clinical, operational and research purposes. Statistical text mining and natural language processing (NLP) are among the popular techniques used in the field\(^5,6\). However, the unstructured, heterogeneous, and context-dependent nature of the clinical documents has made these techniques less effective\(^6,7\). Recent efforts in the development of a section identifier and labeler for preprocessing clinical documents before applying NLP enhances their effectiveness\(^6,8,9,10\).

The section identifier and labeler OBSeCAn, which stands for Object-Based Section Annotator, was developed to preprocess the vast amount of clinical notes stored by the U.S. Department of Veterans Affairs (VA) in the Veterans Health Information Systems and Technology Architecture (VistA) network for various tasks including extracting symptoms and concepts related to specific domains of interest such as homelessness and indwelling urinary catheters. VistA clinical notes are semi-structured documents generated by entering text into pre-defined or user-defined templates\(^8\). The lack of consistency in the structure of the notes makes it difficult to solely use regular expressions to identify sections since these documents do not adhere to particular grammatical rules. OBSeCAn identifies sections and their hierarchical relationships using an ontology that describes the relevant concepts of clinical notes sections and the relationships among these concepts. OBSeCAn identifies labeled and unlabeled sections (sections with and without section headers) in semi-structured clinical documents and annotates these sections to be used for further information extraction tasks.

In a separate study, OBSeCAn was trained and evaluated on a small corpus of 1000 VistA notes including both inpatient and outpatient encounters from any medical specialty. The results showed reasonable metrics for recovering both labeled and unlabeled sections from semi-structured electronic medical notes. In this study, we use all 1000 notes to further train OBSeCAn and evaluate the performance of this method on annotating sections in one million medical documents derived from routine care offered to Veterans at VA medical facilities.

Background

This importance of and motivation for standardization of section headers in clinical notes has been addressed by many authors\(^8,8,9\). This standardization is a key step toward the creation of an ontology for clinical document sections. The existence of such an ontology may be useful for integrating and sharing clinical information among health care professionals and information systems. So far, a terminology of discovered section headers was developed for use with SecTag, a method for automated section header identification\(^8\). The SecTag method was the
first terminology-based method for identifying and labeling sections. In addition, other methods have identified and/or labeled section headers using regular expressions, rules, and machine learnings. Until now, the SecTag method and its study were the best developed and evaluated of the methods for automated section header identification. However, the development of SecTag was based on notes within the Vanderbilt EMR system and demonstrated some limitations when applied to identify and label sections of VA notes. Seeing the potential of terminology-based approach in identifying and labeling sections in the semi-structured notes within VA, OBSecAn was developed using a similar approach to that of SecTag with modifications to the section header terminology, the annotated structure, and the algorithm to provide an automated section annotator for VA notes. In this paper, we discuss the development of the OBSecAn method and its scale to VA big data.

Methods

The OBSecAn algorithm for section annotation identifies the sections and labels the unlabeled sections within semi-structured clinical documents. The components of the OBSecAn method include:

Clinical Document Sections Ontology (CDSO)

A section heading terminology was developed to provide a list of concepts and synonyms for clinical section headings and subsections. The resulting section heading terminology contained 1109 concepts with 4332 synonyms and the hierarchical relationships. As mentioned in Denny et al., the existence of section heading terminology helps to better understand clinical notes. However, we think it is important to not only standardize section headers but also standardize sections. The creation of a clinical document sections ontology (CDSO) is a step toward the standardization of sections for health data integration. The existence of a CDSO can also aid NLP and data mining tasks. For the OBSecAn method, the CDSO helps to recover unlabeled sections. For example, “regular rate and rhythm” is often used under the “Physical Exam” section of a clinical H&P note to describe a result obtained from examining the heart. In other notes the words “regular rate and rhythm” are placed under the section “heart exam” and many times the section headers for the “heart exam” section are omitted. If the phrase “regular rate and rhythm” is added into the CDSO and related to the “heart exam” section concept, it increases the chance of recognizing the “heart exam” section if it is not labelled. For the OBSecAn method, we created a very preliminary CDSO which was evolved from the section header terminology to aid the OBSecAn algorithm. Currently, we have three different types of elements within the CDSO: Section, Section Header, and Property.

A section as defined in Denny et al. is a clinically meaningful grouping of symptoms, history, findings, results, or clinical reasoning that is not itself part of the unique narrative for a patient. In the CDSO, a Section element represents a unit often being used by health care providers to group several elements that have something in common. The way to group the elements in clinical notes varies among health care providers. The CDSO aims to provide an exhaustive list of these groupings. Physicians commonly group into one section several separable sections that can be found in other notes. For example, the section labeled “personal and social history” is a composition of the sections “personal history” and “social history”. We excluded from the CDSO those sections which can be decomposed into more than one section. The structure of a Section element in the CDSO consists of a concept identification, a description, and a data type.

A Section Header element in the CDSO is a found header for a Section element. There are variations of headers being used to label the same Section element across clinical notes. There are also cases where the same header string can be used to label different Section elements. For example, the header string “Cardiac” is used to label the section “Cardiovascular Exam” as well as the section “Cardiovascular Review.” For the header string “Cardiac”, we need to know the context of the document in order to determine which clinical section is being labeled. If this header string is found under the section “Physical Exam” then it is referring to “Cardiovascular Exam”; it refers to “Cardiovascular Review” if placed under the section “Review of Systems.” To assist the OBSecAn algorithm in efficiently labeling the sections, we document in the CDSO aheaders with direct reference to the section like “Cardiovascular Exam” as explicit headers. The CDSO inherited a subset of header strings found in the section header terminology of Denny et al.. In order to achieve good coverage of the section headers found in various clinical notes, we extracted and reviewed headers that appeared frequently in more than 50,000 VA boiler-plate templates generated from all stations within the VA for additional document section headers.

A Property element in the CDSO is a term that often appears in the content of a section. One example is the phrase “regular rate and rhythm”. As mentioned earlier, Property elements in the CDSO play an important role in recovering the unlabeled sections. To generate the list of possible Property elements for a section, we use the content of the labeled sections to obtain a list of frequently found terms used in that section.
There are different types of relationships among the elements of CDSO. The parent-child relationships are added to the CDSO to represent the hierarchical structure of section concepts. The relationships among the concepts are polyheirarchical parent-child relationships. Children of a document section concept are all the subsections found for that section across the clinical notes. A child section may be a part of several different parent sections. CDSO also contains the is-Header relationships between Section Header elements and Section elements and the is-Property relationships between Section elements and Property elements. The relationships among the elements of CDSO were obtained from training on a sample of 1000 documents extracted from VistA notes.

**Section Annotation Output.** With the aim of using OBSecAn to preprocess clinical notes to be used for further focused information extraction, we saved the section annotation output outside of the original document. An example of the format for the section annotation output, viewable with Visual Tagging Tool (VTT) is shown in Figure 1.

For annotating sections in large scale, we simplify the output to contain only the information that is useful for future retrievals of the section content. Once the OBSecAn has annotated sections in a clinical note, the identification of the original document and the annotated section information are saved for future retrieval. The annotated section information includes the section identifications, the section beginning offsets, and the section ending offsets. The section beginning offsets are set to the beginning section content offsets, not the beginning section label offsets.

**OBSecAn Algorithm.** The OBSecAn algorithm has three major stages. The first stage reads and parses the document to obtain and store information regarding sections into a structure that supports the hierarchy of sections. The second stage detects and makes correction to errors in the parsed structure. The third stage produces the section annotation output using the final parsed tree.

The first stage of the algorithm involves processing the documents by sequential chunks of text and storing the processed information into a structure that supports the hierarchy of sections. A clinical note may contain a list of sections and a section may contain a list of nested sections. OBSecAn identifies sections from the note and saves the structure of those identified sections into a supported data structure. This data structure contains a list of nodes representing outermost sections called root nodes. Each node in the list again links to a list of nodes representing the nested sections. OBSecAn sequentially processes a note and creates nodes and inserts them into the data structure. An illustration of data structure used by OBSecAn to store the information of identified section is shown in Figure 2.

**Figure 1.** An example of section annotation output to be viewed with Visual Tagging Tool (VTT)

**Figure 2.** Illustration of data structure used by OBSecAn to store the information of identified sections
created root node and its linked nodes provide the context for disambiguating the next created section. Let ids₁ be the concept identification of the section in the most recent created root node. The concept identifications of the most recent created nested sections subsequently are ids₂, ids₃, ..., idsₙ. The steps for processing a clinical note are as follows:

**Select Text for Annotation.** The text from a clinical document is annotated by chunks. A chunk of text is formed by reading a line from the document. The next line of the text in the document is read and concatenated to the chunk only if the beginning of the next line is a continuation of the line that has already been read. The resulting chunk of text is then split up into segments using “dot/period (.)” as delimiter. We added an extra step to verify whether the dot/period is not a sentence delimiter but rather a decimal separator or a part of an abbreviation. Each text segment is then selected for annotation. The flow chart for annotating sections in a text segment is shown in Figure 3.

**Map Header Candidate String to Section Concepts.** A header candidate string of a text segment is either the whole text segment or the string that begins the text segment and ends right before a colon, dash, comma, semicolon, backslash, or dot. There may be several header candidate strings for a text segment. We examine the header candidate strings in sequence until we find maps of the current header candidate string to the section headers in the CDSO. As previously mentioned, a header candidate string may also be mapped to the header string of several Section concepts.

**Map the Leading Words to Section Concepts.** A section header may appear in a clinical document as leading words of a text segment. For example, the section header “Concentration” of the following text segment “Concentration is very poor”, which appears under the section “Mental Status Exam”, is not delimited from the rest of the text. In this case, we start from the first word of the text segment and find the maximum-length string which can be mapped to Section Headers in CDSO. This maximum-length string may also be mapped to several Section concepts. We also consider trimming the leading words that are prepositions or articles; for example, trimming the part “On the” in the following text segment “On the examination, temperature of the patient is 97.5.”

**Choose Best-Mapped Section Concept.** For each of the mapped Section concepts, we calculate the distance to the recently created sections ids₁, ids₂, ids₃, ..., idsₙ based on the number of unlabeled sections to be inserted into the current section hierarchy. The distance is the minimum number of sections created if the mapped Section concept is to be inserted into the current section hierarchy. The best-mapped Section Concept is the concept among the mapped Section concepts with the minimum distance to the recently created sections. The list of minimal created sections (referred to as unlabeled sections) is also inserted into the current section hierarchy along with the mapped section. In order to reduce the noise in introducing the labels to the unlabeled sections, we set thresholds for the number of created sections. The threshold is set to 3 if the mapped section is derived from mapping to Section Header elements. Otherwise, it is set to 2 if the mapped section is derived from mapping to Property elements. If the number of the unlabeled sections introduced when inserting a mapped section into the current section hierarchy exceeds the threshold for the given case, we omit the insertion of the mapped section and move forward.

**Find first occurrence of Header Strings and Property Strings of a Section.** Searching for Header Strings and Property Strings embedded in the text segment is used to discover the unlabeled sections. The list of Header Strings and Property String of Section idsₙ includes all Header Strings and Property Strings with the relationships is-Header or is-Property to Section idsₙ and its child Sections.

**Retrieve Mapped Section Concept.** The Section concepts associated with the Header String or Property String which first occurs in the text segment that are retrieved. In the case that there are several concepts associated with a Header String or Property String, we consider the best-mapped Section concept.

**Insert Mapped Section.** After a mapped section has been identified from the text segment, we create a node for the mapped section along with the nodes for unlabeled sections. These nodes are linked to the nodes in the data structure for recently created sections ids₁, ids₂, ids₃, ..., idsₙ or linked to a newly created root node.

**Select remain text for Annotation.** The part of the text segment after the text used to identify the mapped Section concept is then being annotated.
**Determining beginning and ending offsets of sections.** The beginning offset of a section is derived along with the position of the Section Header or Property related to the discovery of the section. For determining the ending offset of a section, we have used two different approaches. The first approach sets the ending offset of a section to the beginning of the next discovered section. The second approach derives the ending offset using machine learning. For the OBSecAn method, we used the first approach in which the ending offset of a section is derived from the beginning offset of the next discovered “same level” section. Though Denny et al. have shown that using machine learning could improve the accuracy in setting the section’s ending offset, it has not been tested on large corpora.

After the entire document has been processed and the data structure used to store the identified section has been built, the second stage of the algorithm proceeds to detect and make correction to the errors found in the parsed data structure. We recursively traversed the parsed data structure from the top nodes and employed the following validation checks:

- Check if there were cases that sections were discovered by mistake in the middle of other sections. As a result, in the parsed structure, the nodes corresponding to the section discovered by mistake are being deleted.
- Validate the content of the tagged section using its data type.

Once the recognizable errors in the parsed data structure had been corrected, we calculated the ending offset of the sections via the information available in the final parsed data structure and generate the OBSecAn’s output.

**Evaluation**

The evaluation of the performance of the OBSecAn method on a small corpus was done in a separate study. In this paper, the scale out of OBSecAn to a VA corpus of one million record was evaluated in three steps: (1) the estimation of disk space required to save the annotated sections vs. the original note corpus; (2) identification of high yield sections from the various note titles wherein a high yield section was determined as the ratio of the
number of an annotated section to the total number of notes; and (3) the rate of accurately identifying a particular section of notes from OBSecAn output. For this study, we have used the family history section as an illustrative example. We processed the one million records and extracted the content of the annotated family history sections. The content of these annotated sections was then written to a spreadsheet. Each row in the spreadsheet is the content of one annotated family history section. We then sorted the spreadsheet so that the rows with similar contents were clustered. These clustered rows made it easier to review in groups.

**Document Corpus**

The VA’s VistA network stores a large number of clinical notes (over 2 billion). The notes are generated from various types of medical encounters across different facilities and medical specialties; therefore, the list of sections and section structures are also varied among different type of notes. To evaluate the performance of OBSecAn on a representative sample of VA notes, we extracted a random corpus of one million notes that are stored and made available for research in a secure VA database: Veterans Informatics and Computing Infrastructure (VINCI)\(^4\). These notes organized into approximately 2819 VA Enterprise Note Titles that are further clustered into \(\sim 220\) higher level Note Titles. The million corpus represents a proportional random sample of documents based on the frequencies of the Enterprise Note titles in VINCI.

**Results**

About 75% of the notes in the million document corpus are accounted for by the top 57 Note Titles presented in the corpus (Table 1).

**Table 1.** Frequency distribution of top 57 Note Titles presented in the one-million note corpus.

<table>
<thead>
<tr>
<th>Note Titles</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nursing</td>
<td>23.53</td>
</tr>
<tr>
<td>Primary Care</td>
<td>15.36</td>
</tr>
<tr>
<td>Mental</td>
<td>5.00</td>
</tr>
<tr>
<td>Pharmac</td>
<td>2.45</td>
</tr>
<tr>
<td>Internal</td>
<td>1.88</td>
</tr>
<tr>
<td>Administrative Note</td>
<td>1.84</td>
</tr>
<tr>
<td>Inpatient</td>
<td>1.75</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>1.37</td>
</tr>
<tr>
<td>Social Work</td>
<td>1.31</td>
</tr>
<tr>
<td>Surgery</td>
<td>1.19</td>
</tr>
<tr>
<td>Dental</td>
<td>1.09</td>
</tr>
<tr>
<td>Physical Therapy</td>
<td>1.05</td>
</tr>
<tr>
<td>Podiatry</td>
<td>1.05</td>
</tr>
<tr>
<td>Discharge Summary</td>
<td>0.99</td>
</tr>
<tr>
<td>Emergency</td>
<td>0.92</td>
</tr>
<tr>
<td>Eye</td>
<td>0.90</td>
</tr>
<tr>
<td>Respiratory Therapy</td>
<td>0.87</td>
</tr>
<tr>
<td>Audiology</td>
<td>0.80</td>
</tr>
<tr>
<td>SATP</td>
<td>0.71</td>
</tr>
<tr>
<td>Physical Medicine Rehab</td>
<td>0.65</td>
</tr>
<tr>
<td>Nutrition</td>
<td>0.58</td>
</tr>
<tr>
<td>Optometry</td>
<td>0.56</td>
</tr>
<tr>
<td>Psychology</td>
<td>0.56</td>
</tr>
<tr>
<td>Urology</td>
<td>0.45</td>
</tr>
<tr>
<td>Orthopedic</td>
<td>0.45</td>
</tr>
<tr>
<td>Recreation Therapy</td>
<td>0.45</td>
</tr>
<tr>
<td>Dermatology</td>
<td>0.43</td>
</tr>
<tr>
<td>Occupational</td>
<td>0.41</td>
</tr>
<tr>
<td>Telephone Encounter</td>
<td>0.37</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>0.36</td>
</tr>
<tr>
<td>Hematology Oncology</td>
<td>0.34</td>
</tr>
<tr>
<td>Cardiology</td>
<td>0.32</td>
</tr>
<tr>
<td>Immunization</td>
<td>0.30</td>
</tr>
<tr>
<td>Respiratory</td>
<td>0.28</td>
</tr>
<tr>
<td>Preventive Medicine</td>
<td>0.24</td>
</tr>
<tr>
<td>Critical Care</td>
<td>0.19</td>
</tr>
<tr>
<td>Urgent Care</td>
<td>0.18</td>
</tr>
<tr>
<td>Nurse Practitioner</td>
<td>0.17</td>
</tr>
<tr>
<td>Anesthesia</td>
<td>0.17</td>
</tr>
<tr>
<td>Phys</td>
<td>0.17</td>
</tr>
<tr>
<td>Nephrology</td>
<td>0.15</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.13</td>
</tr>
<tr>
<td>Specialty Care</td>
<td>0.13</td>
</tr>
<tr>
<td>Neurology</td>
<td>0.11</td>
</tr>
<tr>
<td>Procedure</td>
<td>0.11</td>
</tr>
<tr>
<td>Consc</td>
<td>0.11</td>
</tr>
<tr>
<td>C&amp;P</td>
<td>0.10</td>
</tr>
<tr>
<td>Team</td>
<td>0.10</td>
</tr>
<tr>
<td>ENT</td>
<td>0.10</td>
</tr>
<tr>
<td>HBPC</td>
<td>0.09</td>
</tr>
<tr>
<td>Pain</td>
<td>0.09</td>
</tr>
<tr>
<td>Pulmonary</td>
<td>0.09</td>
</tr>
<tr>
<td>Research</td>
<td>0.08</td>
</tr>
<tr>
<td>Mental</td>
<td>0.08</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>0.08</td>
</tr>
<tr>
<td>Individual</td>
<td>0.08</td>
</tr>
<tr>
<td>Primary</td>
<td>0.07</td>
</tr>
</tbody>
</table>

**Annotating and storing sections using OBSecAn**

The original document’s identification, the section annotation output including the identification, and the beginning and ending offsets of the annotated sections were saved for use with further information extraction tasks.

For efficiency in storing and querying data, we divided the one million notes into 20 sets of fifty-thousand notes. All fifty-thousand notes in each set were concatenated and saved into one large file. A note in this large file contains the note identification, the content of the note, and the end of note delimiter. We opened a file stream and annotated sections for all the notes in each of the large file and stored the annotation output into a corresponding file. An output file included the sections annotated for all fifty-thousand notes. The annotated output for each note in the output file includes the note identification, the section annotated information, and the end of note delimiter. The
execution times and disk spaces for storing the original notes and the annotation output for each of the fifty-thousand sets of notes are shown in Figure 4.

![Figure 4](image_url)

**Figure 4.** Disk space in Kilobytes (Kb) for storing the original medical notes and the section annotation outputs for 20 sets of fifty-thousand notes each.

**High Yield Sections and Corresponding Note Titles**

The frequency of a section that appears in notes varies by Note Titles. For each of the Note Titles, the ratio of the number of an annotated section to the total number of notes was used to determine the high yield sections for the corresponding Note Title. There are sections that may appear more frequently in some note title than others. For example, it would be appropriate to expect a completed family history section in a primary care note as opposed to an administrative note. Similarly, sections pertaining to nursing assessments such as the Braden Scale or skin assessment would be expected more frequently in a nursing note. Thus, depending on the information extraction task and domain knowledge, this method would identify a high yield section for natural language processing. For a representative set of 10 of the above Note Titles, we present a list of selected sections and their frequencies (in parentheses):

**Nursing:** Vital Signs (45007), Preventative Medicine (26874), Hygiene (26261), Mobility (24369), Activity Assessment (19238), Patient Education (14050), Review Of Systems (13510), Chronic Pain History (13242), Active Medications (12862), Physical Exam (12769), Functional Status (12153), Activities Of Daily Living (11514), Skin Reassessment (11484), Mental Status (10537), Pain Assessment (10252), Braden Scale (10218), Nursing Comments (10110), Nutrition (9848), Ambulation Status (9245), Functional Status Prior To Admission (9201), Current Skin Assessment (8556), Dressing (8215), Post Op Care (7960), Lab Data (7615), Chief Complaint (7450), Nursing Activity (7253), Morse Fall Scale (7221), Nursing Assessment (7080)

**Primary care:** Vital Signs (72565), Physical Exam (64143), Assessment (46917), Active Medications (44390), Pain Exam (43442), Preventative Medicine (39120), Review Of Systems (36975), Lab Data (32048), Hx Present Illness (30968), Chief Complaint (25914), Past Medical History (17748), Family History (7022), Allergies (12388), Alcohol Audit Screening (10330), Screen For Depression (8176), Clinical Impression (6906), Colorectal Cancer Screen (6361), Reason For Visit (6306), Personal Habits (6010),

**Mental health:** Assessment (13992), Mental Status (7417), Axis I (5524), Axis II (5170), Discharge Diagnosis (4548), Axis III (4367), Axis V (3897), Axis IV (3885), Active Medications (3578), Preventative Medicine (2812), Clinical Impression (2163), Family History (1201), Goal (1040), Gaf (1028), Mental Health/psychiatric (903), Substance Abuse Hx (808), Past Medical History (797), Diagnostic Impression (791), Hx Present Illness (722),

**Inpatient:** Physical Exam (10620), Vital Signs (10128), Assessment (8869), Active Medications (5274), Active Problems (3323), Allergies (3310), Admitting Diagnosis (3224), Hx Present Illness (3131), Family History (747), Chief Complaint (2913), Vitals Interpretation (2894)
**Psychiatry:** Assessment (6381), Mental Status (4373), Axis I (3156), Axis III (3021), Axis II (3004), Axis V (2969), Axis IV (2927), Clinical Impression (2738), Discharge Diagnosis (2387), Active Medications (1920), Preventative Medicine (1433), Active Problems (1021), Additional Diagnosis (867), Family History (344), Gaf (591), Past Medical History (587), Hx Present Illness (581), Lab Data (521), Mental Health/psychiatric (333)

**Surgery:** Preoperative Diagnosis (2812), Procedure (2763), Postoperative Diagnosis (2099), Vital Signs (1692), Review Of Systems (1472), Active Meds (1145), Physical Exam (1095), Reason For Procedure (923), Past Medical History (483), Active Problems (264), Discharge Medications (248), Informed Consent (245), Advance Directive (241)

**Dental:** Diagnosis (8374), Treatment Status (3358), Problem (3254), Past Medical History (393), Medications (539), Chief Complaint (292), Previous Treatments (263), Active Problems (244)

**Physical Therapy:** Physical Exam (979), Functional Status (928), Range Of Motion Exam (811), Past Medical History (677), Statement Of Goal (584), Bed Mobility (579), Procedure (543)

**Podiatry:** Physical Exam (6853), Past Medical History (2345), Musc Exam (2254), Lymp Nodes Exam (1770), Muscle Strength (1669), Range Of Motion (1636), Previous Treatments (1570), Medications (773), Active Problems (567)

**Discharge Summary:** Assessment (7932), Course In Hospital (7255), Hx Present Illness (6076), Physical Exam (5997), Discharge Medications (5070), Discharge Diagnosis (4772), Past Medical History (3996)

**Table 2.** Family history annotations by OBSecAn on the million document corpus. Annotated sections refers to those identified by OBSecAn from the respective documents. Labeled refers to those sections that were designated as “Family History” in the note.

<table>
<thead>
<tr>
<th>Note Titles</th>
<th>Number of Annotated Sections</th>
<th>Labeled Sections (% of total annotations)</th>
<th>Un-labeled Sections</th>
<th>True Family History Annotation (% of total annotations)</th>
<th>Number of Incorrect Annotation</th>
<th>Incorrect due to the Ending Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Care</td>
<td>7022</td>
<td>6753 (96)</td>
<td>269</td>
<td>6916 (98)</td>
<td>106</td>
<td>52</td>
</tr>
<tr>
<td>Nursing</td>
<td>4384</td>
<td>4315 (98)</td>
<td>69</td>
<td>4306 (98)</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td>Mental Health</td>
<td>1201</td>
<td>1190 (99)</td>
<td>11</td>
<td>1177 (98)</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>677</td>
<td>664 (98)</td>
<td>13</td>
<td>659 (97)</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>Int Medicine</td>
<td>536</td>
<td>530 (99)</td>
<td>6</td>
<td>536 (100)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Administrative Note</td>
<td>603</td>
<td>597 (99)</td>
<td>6</td>
<td>596 (99)</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Inpatient</td>
<td>747</td>
<td>743 (99)</td>
<td>4</td>
<td>726 (97)</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>344</td>
<td>335 (97)</td>
<td>9</td>
<td>338 (98)</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Social Work</td>
<td>390</td>
<td>384 (98)</td>
<td>6</td>
<td>387 (99)</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Surgery</td>
<td>307</td>
<td>302 (99)</td>
<td>5</td>
<td>279 (91)</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>Dental</td>
<td>115</td>
<td>111 (97)</td>
<td>4</td>
<td>112 (97)</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Physical Therapy</td>
<td>236</td>
<td>235 (100)</td>
<td>1</td>
<td>236 (100)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Podiatry</td>
<td>219</td>
<td>214 (98)</td>
<td>5</td>
<td>211 (96)</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Discharge</td>
<td>660</td>
<td>657 (100)</td>
<td>3</td>
<td>648 (98)</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Emergency</td>
<td>291</td>
<td>289 (99)</td>
<td>2</td>
<td>285 (98)</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Eye</td>
<td>381</td>
<td>380 (100)</td>
<td>1</td>
<td>381 (100)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Respiratory</td>
<td>129</td>
<td>127 (98)</td>
<td>2</td>
<td>127 (98)</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Audiology</td>
<td>226</td>
<td>222 (98)</td>
<td>4</td>
<td>219 (97)</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>SATP</td>
<td>234</td>
<td>231 (99)</td>
<td>3</td>
<td>233 (100)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Physical Medicine</td>
<td>171</td>
<td>170 (99)</td>
<td>1</td>
<td>166 (97)</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
Accuracy of OBSecAn Annotation of Family History Section

As shown in Table 2, the family history section is present in many note titles in varying frequencies. Primary care, nursing, and mental health are the three note titles with the highest yield for family history sections. It was interesting to note that the family history section was present in nearly half the note titles. This section was labeled as such in almost all instances in all note titles. The rate of accurately identifying the family history as such was between 93% and 100% with a median of 99% in the top 20 note titles shown in Table 2. Most of the inaccuracies were due to an incorrect ending offset of the section.

Discussion

We demonstrate the feasibility of accurately identifying a specific section in VA electronic medical notes, using OBSecAn, an automated ontology-based section annotator. OBSecAn was first trained and tested on a small corpus of documents and now has been scaled out to a million documents from the VA.

Several methods have been developed to sectionize a document. Extending current methods to include a robust clinical documents section ontology and a logical sequence of steps to parse, correct, and ultimately annotate a section based on the final parsed tree are contributions of OBSecAn to the field of automated information extraction. The computational intensity of this algorithm, while significant, is feasible and practical. The stored annotations represent only a fraction of the storage needed for the original documents and thus represents economy of scale while dealing with large document corpora.

The identification of ‘high yield’ sections from specific document note titles represents another conceptual innovation that has the potential to increase the efficiency of information extraction tasks. As noted above, the family history section is present in higher frequency in certain notes, such as primary care. While this is intuitive based on domain knowledge and workflow of clinicians, it offers an objective metric to identify high yield notes based on the specific task and use case. With the example of family history, it was interesting to note that the sections were appropriately labeled as family history in nearly all the top documents identified by their frequency count of the presence of family history sections. The accuracy with which OBSecAn was able to identify a section as family history was nearly perfect in most of the high yield documents reviewed.

We acknowledge several limitations. Even though the million document corpus was representative of the entire VA document corpus, it is possible that there are variations in the documents with respect to their semi-structured format and section beginning and ending offsets. Thus, it is important to further evaluate the sectionizer on a larger scale out. Rather than brute force application of OBSecAn to a larger corpus, we are currently preparing to apply the method to a large corpus and specific use cases such as symptom extraction, homelessness, and detecting the presence of indwelling urinary catheters in hospitalized patients. Accurately identifying a section achieves only one objective. The information extraction task based on specific use cases (such as the family history example discussed above) would need to be evaluated for the relevance of the information extracted. Based on pilot reviews of the textual content of the family history section of several hundred documents, it appears there are several types of templates and free text modalities for recording the family history. Thus, extracting information from this section is a major challenge. We developed this method exclusively on VA documents and, therefore, it may not perform as well in other settings. However, we are interested in extending and applying OBSecAn to electronic notes from other healthcare systems with electronic medical records.

Conclusion

There is a growing recognition of the importance of identifying and labeling sections in clinical notes. The main purpose of the task is to provide the document context for efficient and accurate information extraction and retrieval. In this paper, we have improved on the original implementation of the OBSecAn for automated section identification on a large corpus of electronic medical notes. To the best of our knowledge, this is the first ever attempt to apply and evaluate a section annotator on a large scale.

Acknowledgements: This work was supported by U.S. Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Health Services Research and Development Projects HIR 10-001 (PI: Samore), HIR-10-002 (PI: Gundlapalli), and U.S. National Library of Medicine, National Institutes of Health training grant T15LM007124 (Tran). We would like to express our gratitude to the administration and staff of the VA Informatics and Computing Infrastructure (VINCI) for their support of our project. We also acknowledge...
the staff, resources, and facilities of the VA Salt Lake City IDEAS Center for providing a rich and stimulating environment for NLP research.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent the views of the U.S. Department of Veterans Affairs or the United States Government.

References

Completing Death Certificates from an EMR: Analysis of a Novel Public-Private Partnership
Jacob S. Tripp, PhD¹, Jeffrey D. Duncan, MS², Leisa Finch², Stanley M. Huff, MD¹,³
¹Homer Warner Center for Informatics Research, Intermountain Healthcare, Salt Lake City, Utah
²Utah Department of Health, Salt Lake City, Utah
³Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

With the objective of increasing electronic death registration, Intermountain Healthcare and the Utah Office of Vital Records and Statistics have developed a system enabling death certification from within Intermountain’s electronic medical record (EMR), consisting of an EMR module and an HL7 interface. Comparison of post-intervention death certification at Intermountain Healthcare against a baseline study found a slight increase in the percentage of deaths certified electronically (73% pre vs. 77% post). Analysis of deaths certified using the EMR-module found that they were completed significantly sooner than those certified on paper or using the state’s web-based electronic death registration system (EDRS) (Mean time: Paper = 114.72 hours, EDRS = 81.84 hours, EMR = 43.92 hours; \( p < 0.0001 \)). EMR-certified deaths also contained significantly more causes of deaths than either alternative method (Mean number of causes: Paper = 3.9 causes, EDRS = 4.0 causes, EMR = 5.5 causes; \( p < 0.0001 \)).

Introduction

Death certificates are the primary source of death information throughout the United States. Each state maintains a central death registration system for deaths occurring in their state. As part of maintaining these registries, they institute systems to collect, validate, and archive death records. Frequently, funeral directors initiate the process of creating a new death certificate and then seek cause of death data from medical professionals, including physicians and medical examiners. A death certificate must be completed before a funeral director can proceed with the disposition (burial, cremation, etc.) of the decedent (deceased person).

Over recent years, a majority of state health departments have implemented web-based electronic death registration systems (EDRS) for the gathering and reporting of Death Certificate data.¹ EDRS have several advantages over traditional paper-based systems. Electronic systems are more efficient than paper-based systems, producing higher quality data in a more timely manner.² Paper-based systems require funeral directors to mail or hand-deliver partially completed death certificates to physician certifiers. Once completed, these paper death certificates need to be delivered, in person or by mail to local or state registrars for finalization, including typing the handwritten data into the state’s electronic death records. Electronic systems can reduce or eliminate travel costs incurred by funeral directors in tracking down physicians to complete a death certificate. It is also less likely that waiting for a physician to complete their portion of the certificate will cause delays in the disposition of a decedent. Further, entering data electronically reduces the risk of handwriting-related errors. Government registrars also save time and resources by not having to enter handwritten data into electronic databases.

In 2006, The Utah State Office of Vital Records & Statistics (OVRS) implemented and deployed an EDRS named the Electronic Death Entry Network (EDEN). Utah was the 13th state to deploy an EDRS. EDEN is available to funeral directors, physicians, and medical examiners, as well as health department officials via the World Wide Web. Using EDEN, they can enter and review data from anywhere with an Internet connection.

Problem

EDEN quickly gained nearly universal adoption among funeral directors, but was not as universally adopted among physicians. While nearly all of Utah’s funeral directors are registered users of the EDEN system, in 2010, only 591 of the nearly 5,000 Utah physicians were registered users of the EDEN system. As of March 2010, the funeral director’s portion of the death certificate was completed electronically in EDEN for virtually all death certificates. Conversely, the physician’s portion was being completed electronically for between 60% and 65% of death certificates.
It is not surprising that funeral directors have a higher EDRS adoption rate than physicians. Funeral directors have a vested interest in seeing an increase in the proportion of death certificates completed electronically. Most of the benefits gained by moving from paper death certificates to electronic death certificates accrue to funeral directors and government registrars. On the other hand, aside from the increased flexibility of being able to enter cause of death information without having to receive the paper copy, physicians do not have much to gain by switching to an electronic process. Completing a death certificate using the EDEN EDRS is more complicated for physicians than completing a paper death certificate. Physicians must set up an account on the EDRS in order to use EDEN, and then remember their login name and password in order to access it. The original EDEN user interface lacked clear instructions for novice users. All users, physicians and funeral directors alike were presented with the same screens for data entry, though each user type only needs to enter data into a subset of the available screens. It was not immediately clear to physicians how to complete their portion of the death record. From the physician’s perspective, completing a paper death certificate is much simpler. The confusion stemming from the extraneous (from the physician’s perspective) fields and screens in EDEN was one barrier to physician adoption of the EDEN EDRS.

When a physician certifier chooses not to complete a death certificate electronically, the funeral director must “drop to paper” – print out the partially completed certificate and physically deliver it to the physician for completion (Figure 1). Utah state law requires that deaths be “certified” (cause of death information completed by a licensed provider) within 72 hours of a decedent’s death, and that the death be registered (meaning the death certificate has been fully completed), within 5 days of the time of death. During the first year and a half that EDEN was in operation, 61% of cause of death data entered electronically within the EDEN system was entered within 72 hours of the time of death. When the funeral director had to use the “drop to paper” option, only 34% were certified within 72 hours.

![Figure 1. Pre-intervention death certificate workflow. FD = Funeral Director. MD = Physician. LHD = Local Health Department.](image)

Intermountain Healthcare (Intermountain) is a large integrated delivery network, centralized in the state of Utah, that provides care to approximately half the population of Utah. For death certification, Intermountain has seen EDRS adoption rates among its provider network similar to those seen statewide. During the period from January 2007 through October 2009, Intermountain physicians certified 39% of the 35,093 deaths registered in the state of Utah, 73% using the EDEN EDRS.
Many Intermountain physicians do use the EDEN EDRS for death registration. EDEN adoption rates are particularly high among physicians who complete a high number of death certificates. During the period examined, Intermountain physicians who completed an average of one or more death certificates per month used the EDEN system for 98% of the deaths they certified. This group of “high volume” death certifiers included only 3% of Intermountain physicians who completed at least one death certificate during that period, but accounted for 65% of all deaths certified by Intermountain physicians during that same period.

The vast majority of Intermountain physicians fell into the much larger group of “low volume” death certifiers, certifying less than one death per month on average. These physicians only used the EDEN system for 28% of the deaths that they certified. The large number of “low volume” certifiers collectively represents a “long tail” that accounted for 35% of death certificates completed by Intermountain physicians between January 2007 and October 2009, but only 13.5% of death certificates certified electronically (Table 1). Nearly three-fourths (74%) of the physicians who certified at least one death during the study period certified less than 1 death per year, and these “very low volume” certifiers were even less likely to use the EDEN system, completing only 12% of their death certificates electronically.

<table>
<thead>
<tr>
<th>Death certificates/month</th>
<th>% of all death certifiers</th>
<th>% of all death certificates</th>
<th>% certified electronically (EDEN EDRS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1</td>
<td>96.5%</td>
<td>35.4%</td>
<td>28.0%</td>
</tr>
<tr>
<td>1-2</td>
<td>1.8%</td>
<td>11.7%</td>
<td>89.8%</td>
</tr>
<tr>
<td>2-3</td>
<td>0.4%</td>
<td>4.7%</td>
<td>98.6%</td>
</tr>
<tr>
<td>3-4</td>
<td>0.4%</td>
<td>6.6%</td>
<td>99.6%</td>
</tr>
<tr>
<td>4-5</td>
<td>0.1%</td>
<td>1.1%</td>
<td>99.3%</td>
</tr>
<tr>
<td>&gt;5</td>
<td>0.7%</td>
<td>40.5%</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

Table 1 - Use of EDRS by Intermountain physicians grouped by number of deaths certified per month; Baseline study - January 2007 to October 2009.

Solution

With the objective of increasing electronic death certification, OVRS applied for and received a grant from the Centers for Disease Control and Prevention (CDC) to implement an electronic cause of death interface between the EDEN system and the HELP2 electronic medical record (EMR) system in place at all Intermountain hospitals and clinics. Intermountain has a history of collaborating with the Utah Department of Health (UDOH), especially for electronic data exchanges. The development of this death certificate interface between Intermountain and UDOH is the first instance in the country of death certification done from within an EMR.

This system, that allows deaths to be certified from the EMR, consists of two key components: an EMR module that allows physician users to input the required cause of death data for death certification, and an HL7 interface that allows these data to be transmitted from the EMR to OVRS. Both of these components were developed following significant discussion and collaboration between technical, clinical, and administrative personnel at Intermountain and OVRS, as well as input from the Utah Funeral Directors Association.

The user-facing EMR module was designed in a manner consistent with National Center for Health Statistics (NCHS) standards. The module allows providers to enter cause of death data for a decedent without needing to directly access the EDEN system, eliminating the need for them to remember a separate login and password. As this module was built specifically to be used by physician certifiers, it only includes data fields for the physician’s portion of the death certificate, eliminating some of the confusion physicians had experienced when using OVRS’s EDEN EDRS. (Figure 2)

In addition to simplifying and streamlining the physician certifiers’ data entry process, the new EMR module provides users with access to detailed instructions, produced by the CDC, for properly completing the cause of death section of the death certificate, to support especially those physicians who only certify deaths infrequently. Because the module is integrated into the EMR, it is able to automatically populate required demographic fields, including name, date of birth, gender, and social security number. Stored certifier user data is also automatically entered into the required fields for certifier name, license number, and mailing address. Another advantage of completing a
death certificate from within the EMR is that certifiers can easily switch back and forth between the death certificate module and other portions of the decedent’s medical record (progress notes, laboratory data, medication history, etc.) to gather relevant data for completing the cause of death portion of the death certificate.

**Figure 2 - Demographics and Cause of Death sections of EMR-based death certificate module**

In order to transmit the cause of death data from Intermountain to OVRS, an HL7 interface for cause of death data was created. This HL7 interface supports three distinct types of messages: (1) an ADT^A04 (Patient registration) message is used when sending a new cause of death record to OVRS, (2) an ADT^A08 (Update patient information) message is used to send an update to an already submitted record, and (3) an ADT^A23 (Delete patient encounter) message is used to notify OVRS that a record was submitted in error and should be deleted. Delete messages do not trigger an automatic deletion of a record, as it was determined that personnel at OVRS would like to manually follow up on those instances before a cause of death record is deleted.

The initial HL7 messages were a version 2.4 format agreed upon by engineers at OVRS and Intermountain. As this project grew in national prominence, leaders at the National Center for Vital and Health Statistics (NCVHS) proposed that a standard HL7 message format be developed for use in sending death certificate records from EMRs to electronic vital records registration systems. That proposal led to the development of an HL7 draft standard for trial use that uses HL7 version 2.5.1 messages. Engineers at Intermountain and UDOH OVRS helped to define and vet this standard, and have updated our interface from the original version 2.4 messages to the new standard.
Once these data have been received, they are merged with the funeral director’s portion of the death certificate that continues to be entered into OVRS’s EDEN system. The demographic data in the physician’s portion of the record are compared against records created in EDEN by funeral directors. These demographic data include decedent name, social security number (if applicable), date of birth, gender, and date of death. If a match is found, the two records are merged. In the event of a non-perfect demographic match, records can be reviewed and a merge can be manually triggered by administrative staff at OVRS. OVRS staff report that approximately 40% of the records merge automatically, with the remaining 60% requiring the manual review step.

In addition to transmitting cause of death data to OVRS, Intermountain also stores a copy of the cause of death in the Intermountain Clinical Data Repository (CDR), fulfilling the Meaningful Use Core Measures requirement to record the “Date and preliminary cause of death in the event of mortality in the eligible hospital or CAH.”

The new workflow associated with the EMR-based death certificate module is as follows: When a death occurs, the funeral director and physician can both complete their portions of the record independently and asynchronously, the funeral director using the EDEN EDRS, and the physician using the EMR module. When the funeral director has completed their portion of the record in EDEN, they can check to see if a physician has already completed their portion. If they have not, they contact the physician by telephone or e-mail as before. With the convenience of completing death certificates from within the EMR, some physicians routinely complete their portion of death certificates even before receiving notification from a funeral home. In that scenario, the funeral director now does not need to contact them, and can wait for the two portions of the record to be matched and merged. (Figure 3)

Figure 3. Post-intervention death certificate workflow with electronic interface from Intermountain EMR to EDEN EDRS. FD = Funeral Director. MD = Physician. LHD = Local Health Department.

Despite being fully developed by 2010, due to competing information systems priorities at Intermountain, the death certificate module in Intermountain’s HELP2 EMR only went into pilot production in March of 2011 with pilot users at both inpatient and outpatient sites, and was not available broadly throughout Intermountain Healthcare until July 2013.

Methods
With the module finally available to most Intermountain physicians, we wanted to assess the impact of being able to complete death certificates from within the EMR. Data was gathered for all deaths occurring in 2014 that were registered in the state of Utah, and compared them against the baseline data for deaths occurring between January 2007 and October 2009.

As the EMR-based death certificate tool was only available to physicians affiliated or employed by Intermountain, we limited the analysis of these data to those providers. Presence of a certifier’s license number in Intermountain’s master service provider registry numbers was used as an indicator of whether or not a physician certifier would have been likely to have had access to the EMR-based death certificate module. Each death was classified according to the method by which it was certified, whether it was certified on paper (“Paper”), via direct use of the EDEN EDRS (“EDEN”), or by using the EMR death certificate module and associated HL7 interface (“EMR”).

We wanted to know whether the introduction of the EMR death certificate module would increase of the proportion of deaths certified electronically at Intermountain, and whether it would increase the percentage of “low volume” certifiers certifying deaths electronically, or simply result in certifiers who were already certifying deaths electronically switching from one electronic method (the EDEN EDRS) to another (the EMR-based death certificate module). To assess the impact of the EMR death certificate module on “low volume” certifiers, we looked at the proportion of deaths certified by each method within certifiers groups stratified by the number of deaths they certified in 2014.

To assess the impact of the EMR death certificate module on the “timeliness” of death certification and registration, we calculated what percentage of deaths certified by each method of death certification (“Paper”, “EDEN”, and “EMR”) was certified within 72 hours of the recorded time of death, together with average time of certification for each method. We also calculated the percentage of deaths certified by each of the three methods that was fully registered within the 5 days allowed under state statute, together with the average time of registration for records in each category.

Additionally, to assess whether having the death certificate certified from within the decedent’s medical record (in the EMR) aided certifiers in more completely describing the cause of death, we calculated the average number of ICD-10 codes assigned to each category of death certificates by coders at the National Center for Health Statistics (NCHS).

**Results**

In 2014, 17,267 deaths were registered with the Utah Department of Health’s Office of Vital Records and Statistics. Of these, 13,249 (77%) were certified by 1537 providers whose license numbers were in Intermountain Healthcare’s service provider registry. Of these 13,249 deaths, 3,051 (23.0%) were certified by 971 certifiers using the “drop-to-paper” method and 10,198 (77.0%) were certified electronically by 675 certifiers, either by using the EDEN EDRS system directly, or by using the EMR-based death certificate module. The 77% certified electronically represents a statistically significant increase over the 73% done electronically in the period from January 2007 to October 2009 (p < 0.001, student’s t test). Of the 10,198 deaths certified electronically, 751 were completed by 181 certifiers using the EMR-based death certificate module, 7.4% of electronically-certified deaths, and 5.7% of all Intermountain-certified deaths in 2014.

Stratifying certifiers by the average number of deaths they certified each month (Table 2), we found that the increased proportion of records certified electronically in 2014 was almost entirely due to a nearly 20% increase in the proportion of records certified electronically by the “low volume” certifiers (those certifying less than 1 death per month). This increase included a 9.5% increase in deaths certified in the EDEN EDRS system, and 10% of deaths certified by this group of certifiers being certified using the now available EMR-based death certificate module. The majority of deaths certified by higher volume certifiers continued to be done electronically, though not quite at the near universal rate seen in the baseline analysis.

On average, deaths certified on paper took 4.78 days from the time of death until the time of certification, with 35.5% being certified within the state-mandated 72 hours. Deaths certified via the EDEN EDRS took an average of 3.41 days, with 56.6% certified within 72 hours of death. Deaths certified via the EMR-based death certificate module took an average of 1.83 days, with 77.5% certified within 72 hours of death. (Table 3)
Table 2 - Comparison of death certification methods used, stratified by the frequency of deaths certified, for baseline (2007-2009) and post-intervention (2014) periods.

Death registration (full completion of the death certificate) for deaths certified on paper took an average of 5.38 days, with 58.5% being fully registered within the 5 days specified by state law. Deaths certified in the EDEN EDRS were registered, on average, 4.55 days after the death, with 68.6% being registered within the mandated 5 day period. Deaths certified in the EMR-based death certificate module were registered in an average of 4.18 days from the time of death, with 74.7% being fully registered within 5 days. Deaths certified via the EMR-based death certificate module were certified significantly sooner than those done using the EDEN EDRS (p < 0.0001, student’s t test). Deaths certified in the EMR were also registered significantly sooner than those certified using the EDEN EDRS (p = 0.0086, student’s t test). (Table 3)

Table 3 - Comparison of timeliness of certification (completion of the physician’s portion of the record) and registration (full completion of the record), as grouped by method of certification. *= p-values calculated using student’s t test.

The average number of ICD-10 coded causes of death assigned by NCHS coders to deaths certified on paper was 3.9. The average number of coded causes of death assigned to deaths certified via the EDEN EDRS was 4.0. The average number of coded causes of death assigned to deaths certified using the EMR-based death certificate module was 5.5. The difference between number of coded causes of death for the paper-certified and EDRS-certified deaths was not statistically significant (p = 0.2049, student’s t test), but the difference between the number of coded causes of death for the EMR-certified deaths was significantly higher than both paper-certified deaths (p < 0.0001, student’s t test) and EDRS-certified deaths (p < 0.0001, student’s t test).
Discussion

In the period between 2009 and 2014, the percentage of deaths certified electronically at Intermountain Healthcare increased from 73% to 77%. This increase was statistically significant, but relatively modest in size, still requiring funeral homes to hand deliver nearly 1 of every 4 death certificates to physicians for certification.

The increase in the overall percentage of deaths certified electronically corresponds with increased electronic certification among certifiers who certified less than 1 death certificate per month. All other strata actually saw a decrease in the percentage of deaths certified electronically, as compared with the baseline study, but because the “low volume” certifiers certify such a large percentage of deaths, the increase in electronic certification in that stratum of certification was enough to increase the overall percentage of deaths certified electronically.

The introduction of a death certificate module within Intermountain’s HELP2 EMR may have contributed a small amount to that increase, especially considering that the increased percentage of deaths certified electronically by “low volume” certifiers was shared about equally between the state’s EDRS and Intermountain’s EMR-based module. However, to date, adoption of the EMR-based death certificate module by Intermountain physicians has been relatively low. Only about 12% of Intermountain certifiers used the EMR-based module to complete at least one death certificate.

We examined the mode of death certification in frequency-based strata because of our hypothesis that the EMR-based module might be more user-friendly for providers who only certify death rarely. However, in light of the fact that the number of providers in most of the frequency-based strata changed significantly between the times of the two samples, it is difficult to attribute much of the changes between the sample periods to changes in behavior by individual providers. In the post-implementation sample, all but the lowest volume stratum had many more certifiers than in the pre-implementation sample, so changes in mode of certification in each stratum are more likely to have come from the behavior of different physicians being in those strata than to have come from a change in behavior by individual physicians. For example, in the “high volume” stratum, the percentage of deaths certified on paper increased from 0.1% to 2.8%, but the number of certifiers increased from 13 to 38. It seems more likely that the increase of paper-based certification is due to a few more physicians who typically certify deaths on paper certifying more than five deaths per month in 2014, than it is due to physicians who previously certified deaths electronically switching back to certifying them on paper.

In each of the frequency-based strata, the EDEN EDRS was used more frequently for death certification than the EMR-based module. The EMR-based module was used most frequently by certifiers completing 4 or fewer death certificates each month. In the “high volume” stratum, certifiers used the EDEN EDRS almost exclusively when certifying deaths electronically. The “high volume” certifiers were already certifying nearly all of their deaths using the EDRS prior to the implementation of the EMR-based option, and perhaps the benefits of the EDRS outweigh those of the EMR-based option for high volume certifiers.

The web-based EDEN system can be used anywhere that a browser is available, while the EMR-based death certificate module requires the installation of a bit of software on each computer where it is to be used. This generally limits use of the EMR-based module to only those computers in Intermountain hospitals and clinics. Additionally, during this time the study was done, OVRS introduced a physician-specific portal for the EDEN system (EDEN MD) that simplifies and streamlines the physician’s part of the process, a significant improvement over the original EDEN system. Additionally, use of the EMR death certificate module has not been mandated at Intermountain, and while most physician certifiers have access it, they may not be aware of its availability. Going forward, efforts could be made to advertise the module’s existence, highlighting some of its conveniences. The small, but not insignificant, rate of the module’s use among low volume certifiers shows some promise that further growth could be seen among that group.

It should not be surprising that deaths certified electronically are both certified and registered significantly sooner than deaths certified on paper, but it is interesting that deaths certified using the EMR-based module are certified and registered significantly sooner than those completed using the state’s web-based EDRS. While it is possible that the early adopters of the EMR-based module are inherently more likely to complete deaths sooner than other certifiers, there are several factors that contribute to the increased timeliness of the EMR-certified deaths. First, the certifiers use the EMR regularly as part of their daily medical practice, and often complete other clinical notes.
related to a patient’s death. A physician completing other clinical documentation on a patient death can simply click over to the death certificate module and certify the death as well. Additionally, certifiers using the EDRS must first wait for a funeral home to begin a death record and then notify them that it is ready for them to certify. This notification may not come at a time when it is convenient for the certifier to log into the EDRS and complete the record, increasing the length of time before the death is certified and ultimately registered.

In addition to being completed significantly sooner than EDRS-certified deaths, the EMR-certified deaths contained significantly more information, as represented by the number of causes of death reported, than either EDRS-certified or paper-certified deaths. Certifiers using the EMR-based module can click back and forth between the death certificate module, and other EMR modules, including clinical notes, lab results, vital signs, etc. One physician reported that, when completing a death certificate within the EMR, he often reviews a patient’s history & physical (H&P) note, and the death summary note to better understand the chain of events leading up to a patient’s death. Further analysis is needed to determine whether the increased number of causes of death in the EMR-certified records can be attributed to having convenient access to the decedent’s medical record, but it seems likely that this would at least contribute to the increase in information reported.

It should be noted that this was not a randomized trial of the EMR-based death certificate module, and physician certifiers were at liberty to use or not use the module as they chose, so there is a possibility of selection bias. It may be that those physicians who choose to use the EMR-based module over the state’s EDRS or paper are more likely than other physicians to complete their death certificates quickly, and include more causes of death.

As mentioned above, one outcome of this work has been the creation of an HL7 Draft Standard for Trial Use for death reporting, specifically designed for the transmission of death certification records from an EMR to an EDRS. That standard has been implemented in the interface between Intermountain Healthcare and the Utah Department of Health’s Office of Vital Records and Statistics. Intermountain is currently transitioning from a set of home-grown EMR applications to a widely used commercially-vended EMR system. As part of that transition, the vendor has added a form for death certification to their EMR. This EMR is now in place at a few facilities, and records entered into its death certificate form are subsequently formatted using the HL7 Draft Standard and transmitted to the EDEN EDRS, using the same interface used for death certifications coming from the HELP2 EMR. Other customers of this commercial EMR could potentially use this same death certificate form and message format to send death record electronically to a government-run EDRS. While Utah’s EDEN system is currently the only EDRS supporting this standard, its implementation in one of the more widely used commercial EMR means that other healthcare organizations that use that EMR could potentially begin certification of deaths from within an EMR. These steps open the door to the standard’s implementation in other EDRS, and the possibility of more widespread EMR-based death certification.

Future Directions

Going forward, there are a variety of enhancements that could be made to the EMR-based death certificate module to increase its utility. Plans for future versions of the module include presenting relevant data (labs, progress notes, etc.) to the end user within the module itself, eliminating the need for switching between modules. There continue to be concerns about the accuracy of death certificates11,12, and another possible enhancement to the EMR-based module would be to develop decision support logic that uses existing patient data to make suggestions of conditions that ought to be considered among the causes of death. Additionally, enhancements to the communication aspect of this system could be made, including supporting the capability of funeral directors to trigger death certificate requests to Intermountain physicians via the EMR’s secure internal messaging system.

Conclusions

The partnership between Intermountain Healthcare and the Utah Office of Vital Records and Statistics succeeded in implementing the first ever EMR-EDRS interface for electronic death certification. To date, the EMR-based death certificate module has had limited adoption among Intermountain certifiers, primarily among users certifying 4 or fewer deaths per month. Still, it may have had a modest impact on the significant increase in electronic death certification at Intermountain Healthcare. Deaths certified using this module were completed significantly sooner than deaths certified using the state’s electronic death registration system, and included significant more causes of death.
death, so there is strong incentive to work to increase the use of this method, both within Intermountain Healthcare, and at other sites in Utah, and in other states.

References


A Method to Compare ICF and SNOMED CT for Coverage of U.S. Social Security Administration’s Disability Listing Criteria
Samson W. Tu, MS, Csongor I. Nyulas, Tania Tudorache, Mark A. Musen
Stanford University, Stanford, CA, USA

Abstract
We developed a method to evaluate the extent to which the International Classification of Function, Disability, and Health (ICF) and SNOMED CT cover concepts used in the disability listing criteria of the U.S. Social Security Administration’s “Blue Book.” First we decomposed the criteria into their constituent concepts and relationships. We defined different types of mappings and manually mapped the recognized concepts and relationships to either ICF or SNOMED CT. We defined various metrics for measuring the coverage of each terminology, taking into account the effects of inexact matches and frequency of occurrence. We validated our method by mapping the terms in the disability criteria of Adult Listings, Chapter 12 (Mental Disorders). SNOMED CT dominates ICF in almost all the metrics that we have computed. The method is applicable for determining any terminology’s coverage of eligibility criteria.

Introduction
Under Title II of the Social Security Act, the U.S. Social Security Administration (SSA) provides Disability Insurance (DI) benefits for workers. Under Title XVI of the Act, the SSA administers Supplemental Security Income (SSI) payments for the poor, aged, and disabled with no or little work experience. Each year, SSA renders decisions on millions of cases at the initial and subsequent reconsideration and appeal levels. To determine whether an adult is disabled, SSA’s regulations provide a five-step sequential evaluation process, based on the statutory definition of disability. The third step in the sequential evaluation process for an adult claimant is the “listing level”, where an SSA claim adjudicator determines whether the severity of a claimant’s impairment meets or medically equals a specific listing and whether the impairment meets a duration requirement. The listings are enumerated and organized into organ-system–dependent chapters in SSA’s “Blue Book,” which has separate parts for adult and childhood listings. Figure 1 shows a partial view of the Blue Book’s chapter on adult mental disorders.

At present time the evaluation of disabilities is largely a document-driven process, where only the International Classification of Diseases (ICD) is used to code data relating to diseases. If more structured data documentation can be introduced into the process, it will have enormous impact on the ability to expedite the evaluation process, to enhance the collection of health statistics, and to enable health service research on disabilities. On September 9, 2011, the Health IT Standards Committee’s Clinical Quality Measures Workgroup and the Vocabulary Task Force jointly recommended that, for functional status, International Classification of Functioning, Disability and Health (ICF) be used for coding categories of function, Logical Observation Identifiers Names and Codes (LOINC) be used for assessment instruments, and SNOMED CT be used for appropriate responses [1].

SNOMED CT and LOINC are well-known standard terminologies in biomedical informatics. ICF is a multipurpose classification that, together with ICD, is a reference classification in the WHO Family of International Classifications (WHO-FIC). It provides a standard language and conceptual basis for the definition and measurement of functions and disability. Unlike a medical model of disability, which sees loss of functions only as consequences of diseases and disorders, ICF embodies a “bio-psycho-social synthesis” that conceptualizes function and disability in the context of health conditions, environmental factors, and personal factors. Therefore ICF codes are organized as hierarchies within the ICF components Body Function, Body Structure, Activity and Participation, and Environmental Factors [2].

1 20 C.F.R. §§ 404.1520 and 416.920
2 http://www.ssa.gov/disability/professionals/bluebook/AdultListings.htm
The Blue Book listing criteria, because of its use in making decisions about claimants’ eligibility, plays a central role in defining the type of data that should be captured in the evaluation process. In particular, because several chapters of the Blue Book include criteria that require the assessment of a claimant’s functional capacity, we developed a method to investigate the extent to which ICF and SNOMED CT can be used to code the concepts and relationships appearing in the listings. To derive the concepts and relationships that should be mapped, we decomposed the listing criteria using a structured schema. To take into account inexact matches between listing concepts and terms from ICF and SNOMED CT, we defined different mapping types. We defined various metrics for measuring the coverage of each terminology, taking into account the effects of inexact matches and frequency of occurrence. Finally, we performed a proof-of-concept validation of our method by mapping the terms in the disability criteria of Adult Listings, Chapter 12 (Mental Disorders). In this small-sample study, we found that SNOMED CT dominates ICF in almost all the metrics that we have computed.

**Methods**

The first problem in using ICF to code health and health-related states is that ICF codes (such as “b210 Seeing functions”) denote possible body functions, body structures, activities and participations, and environmental factors that may affect a person’s function. They are not codes for specific conditions that a person may have. These codes need to be combined with *qualifiers* (such as “3 Moderate Impairment”) in order to describe a health-related state. To map disability criteria into ICF codes, it is necessary to decompose descriptions of a claimant’s physical and mental states into a format that facilitates matching ICF’s specific coding scheme. Much has been written on the representation of eligibility criteria [3]. For this coverage study, we adapted our earlier work on annotating human study eligibility criteria [4] to create a *mapping schema* that specifies, for a particular listing criterion, how to identify concepts and relationships in the criterion and how to decompose them in a structured format that makes clear the semantic roles played by these concepts and relationships. For example, the criterion
A residual disease process that has resulted in such marginal adjustment that even a minimal increase in mental demands or change in the environment would be predicted to cause the individual to decompensate can be decomposed into the concept *residual disease process* having the relationship *has result* with another complex concept Y, where Y can be structured as a marginal adjustment that is defined by another complex concept *a minimal increase in mental demands or change in the environment would be predicted to cause the individual to decompensate*. We decompose the concept recursively until we end up with atomic concepts and relationships that cannot be further decomposed. These atomic concepts and relationships are the entries that we want to map to standard terminologies.

More specifically, we define as atomic a *Primitive Concept* (an entity representing an idea that stands by itself and that we don't further decompose (e.g., *disease process*)), a *Qualifier* (a relationship or attribute that can be used to specialize the meaning of a concept; e.g., *result in*), and a *Qualifier value* (an attribute value that describes a quality of the thing represented by the concept; e.g., *high*). Because of our special interest in ICF, we define *Function* (e.g., *short-term memory*) as a special kind of *Primitive Concept*.

We define composite concepts recursively using other concepts and relationships. A *Boolean concept* is made up of a combination of other concepts using Boolean operators *AND*, *OR*, and *NOT*. A *Qualified concept* consists of a root concept and a set of *Qualifiers* and *Qualifier values* that are implicitly combined using *AND*. A *Change in function* concept consists of a concept representing one or a combination of functions and a change descriptor (an instance of *Qualifier value*), such as *impaired* or *significant loss*, that describes how the function has changed. Finally, we have an *Event collection* concept that represents a set of events (e.g., *repeated episodes of decompensation*). An instance of *Event collection* that is associated with a concept (e.g., *decompensation*) represents multiple occurrences of the event signified by the concept.

We define *Mapping* as having two properties: a *Mapped code* and a *Mapping type*. An instance of *Mapped code* principally contains attributes that identify the code and terminology of the matched concept. We defined five different types of mappings, as shown in Table 1.

<table>
<thead>
<tr>
<th>Mapping type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>equivalence</td>
<td>The mapped code is equivalent to the concept, qualifier value, or qualifier being mapped</td>
</tr>
<tr>
<td>generalization</td>
<td>The mapped code is a generalization of the concept, qualifier value, or qualifier being mapped</td>
</tr>
<tr>
<td>specialization</td>
<td>The mapped code is a specialization of the concept, qualifier value, or qualifier being mapped</td>
</tr>
<tr>
<td>partial match</td>
<td>The meanings of the mapped code and the concept, qualifier value, or qualifier being mapped overlap</td>
</tr>
<tr>
<td>related</td>
<td>The mapped code is related to the concept, qualifier value, or qualifier being mapped, although their meanings don't overlap</td>
</tr>
</tbody>
</table>

We modeled the mapping schema in Protégé, an ontology editor developed at Stanford. In addition to the types of concepts and qualifiers, mapping code, and listing criteria, we have classes that represent the ICF chapters and criteria groups in a chapter. The result is a small ontology consisting of the entities (Figure 2) and properties of the mapping schema.

Protégé allows us to model the groupings of the criteria in a Blue Book chapter and to annotate the criteria with concepts, to decompose the concepts, and to map the concepts to codes in terminologies that are available in Bioportal, a repository of biomedical ontologies and terminologies maintained by the National Center for Biomedical Ontologies.

---

3 More information about Protégé can be found at http://protege.stanford.edu

4 http://bioportal.bioontology.org
Figure 2. Class hierarchy representing entities in the mapping schema. Entities in the Annotatable class hierarchy allow us to represent the Listing_criteria of a Listing_chapter as part of Criteria_groups. The Listing_criteria are decomposed into their constituent Primitive_concept, Qualifiers, and Qualifier_values, which have Mappings that map them to Mapped_code from standard terminologies.

A Bioportal reference widget [5] for Protégé allows us to search for terminology codes available in Bioportal from within Protégé (Figure 3). We configured this widget for the has_mapped_code property. A user needing a code for the property can search within the widget codes or terms in specified terminologies or ontologies. Once he or she decides that a particular code is appropriate, clicking the "Import" button automatically creates, in Protégé, the mapped code corresponding to the selected code in the search results.

Once we had defined the mapping schema, two developers independently decomposed the criteria in the Adult Mental Disorder chapter of the Blue Book and mapped the concepts and relationships derived from the decompositions to ICF and SNOMED CT. They reconciled their work to come up with the consensus decompositions and mappings. To allow people not familiar with Protégé to access the mappings, we wrote scripts that export the mapping into a variety of formats, including HTML and text (Figure 4).

Next we define the metrics for quantifying the mapping results. Evaluations of a terminology's coverage of a subject domain usually define coverage as the proportion of terms and relationships harvested from a corpus in the subject domain that can be represented using terms and relationships in the terminology [6-12]. The coverage metrics may be refined by assigning “importance” weights to the concepts to be matched [7], by weighing the quality of matches [12], by the use of post-coordinated concepts to match target terms [10], and by the frequency of occurrences of the concept in the corpus [10]. Because the decompositions of different criteria may contain the same concepts, the leaf instances in our decompositions of listing criteria include duplicate concepts. Furthermore, sections of the Mental Disorders chapter may contain criteria that already occur in other sections. Together with the two weighting schemes
of the mapping types (equal weights or weights of \{1.0, 0.7, 0.7, 0.5, 0.3\} for the 5 mapping types \{equivalence, generalization, specialization, partial match, related\}), we can define 6 metrics as shown in Table 2.

Figure 3. The use of Bioportal reference widget in Protégé to search for 'short term memory' in ICF. The widget matches the search term among all descriptors of entities in the selected ontology and displays the search results (e.g., "b1440. Short-term memory") as rows in a table. To the right of each search result there are clickable buttons to perform operations such as importing the selected term into Protégé ontology.

Class 'Listing criteria'
- **text**: Difficulty concentrating or thinking
- **concept**: Class 'Boolean concept'
  - **text**: (Difficulty concentrating) OR (Difficulty thinking)
  - **boolean operator**: OR
  - **concept**:
    1. Class 'Change in function'
       - **text**: Difficulty concentrating
       - **involved function**: Class 'Function'
         - **text**: Concentrating
         - **mapping**: Class 'Mapping'
           - **mapped code**: Class 'Mapped code'
             - **ontologyId**: ICF
             - **label**: b1400. Sustaining attention
           - **mapping type**: equivalence
         - **change descriptor**: Class 'Qualifier value'
           - **text**: Difficulty with
    2. Class 'Change in function'
       - **text**: Difficulty thinking
       - **involved function**: Class 'Function'
         - **text**: Thinking
         - **mapping**: Class 'Mapping'
           - **mapped code**: Class 'Mapped code'
             - **ontologyId**: ICF
             - **label**: d163. Thinking
           - **mapping type**: equivalence
         - **change descriptor**: Class 'Qualifier value'

Figure 4. A display showing the decomposition and mappings of the criterion "Difficulty concentrating or thinking" in a generated HTML file. The decomposition ends up with two instances of Change in function that have \{Difficulty with, Concentrating\} and \{Difficulty with, Thinking\} as terms that need mapping. ‘Concentration’ is shown to have equivalence match with ICF code ‘b1400. Sustaining attention’ and ‘Thinking’ is shown to have equivalence match with ICF code ‘d163. Thinking.’
Table 2. Definition of metrics used to evaluate the coverage of terminologies.

<table>
<thead>
<tr>
<th>Metric #</th>
<th>Target terms to be mapped</th>
<th>Weights on mapping types {equivalence, generalization, specialization, partial match, related}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Concepts, qualifiers, and qualifier values (with no duplicate)</td>
<td>{1, 1, 1, 1}</td>
</tr>
<tr>
<td>2</td>
<td>Concepts, qualifiers, and qualifier values (with no duplicate)</td>
<td>{1.0, 0.7, 0.7, 0.5, 0.3}</td>
</tr>
<tr>
<td>3</td>
<td>Bag of concepts, qualifiers, and qualifier values that are leaf nodes in the decomposition of distinct Chapter 12 criteria (no repetition of criteria)</td>
<td>{1, 1, 1, 1}</td>
</tr>
<tr>
<td>4</td>
<td>Bag of concepts, qualifiers, and qualifier values that are leaf nodes in the decomposition of distinct Chapter 12 criteria (no repetition of criteria)</td>
<td>{1.0, 0.7, 0.7, 0.5, 0.3}</td>
</tr>
<tr>
<td>5</td>
<td>Bag of concepts, qualifiers, and qualifier values that are leaf nodes in the decomposition of all Chapter 12 criteria (possible repetition of criteria)</td>
<td>{1, 1, 1, 1}</td>
</tr>
<tr>
<td>6</td>
<td>Bag of concepts, qualifiers, and qualifier values that are leaf nodes in the decomposition of all Chapter 12 criteria (possible repetition of criteria)</td>
<td>{1.0, 0.7, 0.7, 0.5, 0.3}</td>
</tr>
</tbody>
</table>

We wrote code to traverse the consensus Protégé data structure of criteria, their decompositions, and mappings to output the criteria, concepts, and mapping information to a spreadsheet. The computation of metrics was implemented in Excel. We ran the code for SNOMED CT mappings and for ICF mappings.

Because ICF is designed to capture results of functional assessments, it necessarily has poor coverage of non-functional concepts. To correct for this bias, we made a tentative assignment of the criteria into functional and non-functional groups, and computed the metrics for each group.

Finally, we disaggregated concepts, qualifiers, and qualifier values and examined the coverage of ICF and SNOMED CT for them separately.

Results

In this proof-of-concept experiment, we found that the Adult Mental Disorders chapter of SSA’s Blue Book has 58 distinct “atomic” criteria that are not defined in terms of other criteria. These criteria reference 114 distinct atomic concepts, 18 distinct qualifiers, and 40 distinct qualifier values. There are numerous groupings of criteria. For example, Generalized persistent anxiety is defined in terms of subcriteria such as Motor Tension and Apprehensive expectations. Generalized persistent anxiety itself is part of a group of criteria used to define Anxiety-related disorders. In the tables that display the results of our mappings, “criteria” refers only to the atomic criteria, not the higher-level criteria that are defined in terms of other subcriteria.

Tables 3 and 4 display the coverage metrics of ICF and SNOMED CT for concepts, qualifiers, and qualifier values (Results for ICF mappings have shaded background in the tables for better readability.) As to be expected, SNOMED CT has a higher percentage of mapped concepts. We can also see that functional criteria are often duplicated in different sections of the chapter. So that when we weigh the metric to include all occurrences of concepts anywhere in the chapter (column 4 of each table), the metrics of ICF coverage for the functional criteria are significantly higher than when we consider only mappings of distinct concepts and qualifiers.

What may be surprising in Tables 3 and 4 is that SNOMED CT also dominates when we examine the functional criteria alone. The relatively low coverage rate of ICF, however, can be explained in part by the low coverage of qualifiers and qualifier values, especially qualifiers. If we compare the mapping rates of concepts only, excluding qualifiers and qualifier values, as shown in Tables 5 and 6, the rates of mappings for functional criteria increase substantially. In fact, if we measure in terms of the all occurrence of concepts in functional criteria, ICF has slightly
higher coverage (86.7%) than SNOMED CT (84.39%). This finding results from relatively higher frequent
occurrences of functional criteria in the Adult Mental Disorders chapter of the SSA’s Blue Book.

We can also see that the reduction in the metrics when we give inexact matches less weight is smaller in the case of
SNOMED CT mappings than that of ICF mappings. This can be explained by the more granular nature of SNOMED
CT concepts, which leads to more exact matches.

Table 3. Coverage of ICF for concepts, qualifiers, and qualifier values in Mental Disorders criteria. For
brevity, "concept" in this table denotes all concepts, qualifiers, and qualifier values.

<table>
<thead>
<tr>
<th>Types of criteria</th>
<th>No concept repetition</th>
<th>No criteria repetition</th>
<th>All occurrences of concept in chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional criteria</td>
<td>48.48%</td>
<td>59.69%</td>
<td>68.63%</td>
</tr>
<tr>
<td>Non-functional criteria</td>
<td>37.97%</td>
<td>40.23%</td>
<td>31.52%</td>
</tr>
<tr>
<td>All criteria</td>
<td>43.82%</td>
<td>51.85%</td>
<td>57.09%</td>
</tr>
</tbody>
</table>

Table 3. Coverage of ICF for concepts, qualifiers, and qualifier values in Mental Disorders criteria. For brevity, "concept" in this table denotes all concepts, qualifiers, and qualifier values.

Table 4. Coverage of SNOMED CT for concepts in Mental Disorders criteria. For brevity, "concept" in this table denotes all concepts, qualifiers, and qualifier values.

<table>
<thead>
<tr>
<th>Types of criteria</th>
<th>No concept repetition</th>
<th>No criteria repetition</th>
<th>All occurrences of concept in chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional criteria</td>
<td>76.47%</td>
<td>80.18%</td>
<td>84.53%</td>
</tr>
<tr>
<td>Non-functional criteria</td>
<td>80.00%</td>
<td>79.45%</td>
<td>83.63%</td>
</tr>
<tr>
<td>All criteria</td>
<td>77.37%</td>
<td>79.89%</td>
<td>84.25%</td>
</tr>
</tbody>
</table>

Table 4. Coverage of SNOMED CT for concepts in Mental Disorders criteria. For brevity, "concept" in this table denotes all concepts, qualifiers, and qualifier values.
Table 5. Coverage of ICF for concepts (excluding qualifiers and qualifier values) in Mental Disorders criteria.

<table>
<thead>
<tr>
<th>Mapping weights equals 1</th>
<th>Types of criteria</th>
<th>No concept repetition</th>
<th>No criteria repetition</th>
<th>All occurrences of concept in chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional criteria</td>
<td>66.67%</td>
<td>76.92%</td>
<td>87.60%</td>
<td></td>
</tr>
<tr>
<td>Non-functional criteria</td>
<td>46.67%</td>
<td>56.14%</td>
<td>47.71%</td>
<td></td>
</tr>
<tr>
<td>All criteria</td>
<td>56.99%</td>
<td>68.15%</td>
<td>75.49%</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Coverage of SNOMED CT for concepts (excluding qualifiers and qualifier values) in Mental Disorders criteria.

<table>
<thead>
<tr>
<th>Mapping weights equals 1</th>
<th>Types of criteria</th>
<th>No concept repetition</th>
<th>No criteria repetition</th>
<th>All occurrences of concept in chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional criteria</td>
<td>85.00%</td>
<td>85.25%</td>
<td>84.39%</td>
<td></td>
</tr>
<tr>
<td>Non-functional criteria</td>
<td>82.61%</td>
<td>83.67%</td>
<td>85.71%</td>
<td></td>
</tr>
<tr>
<td>All criteria</td>
<td>83.17%</td>
<td>84.55%</td>
<td>84.87%</td>
<td></td>
</tr>
</tbody>
</table>

Discussion

We have developed a method for evaluating the extent to which ICF and SNOMED CT cover the concepts and relationships used to describe SSA Blue Book listing criteria. We applied the method to the criteria of the Adult Mental Disorders chapter and computed a number of metrics for comparing the coverage of the two terminologies. Our findings suggest that, compared to ICF, SNOMED CT has broader coverage of both functional and non-functional concepts in a Blue Book chapter that is rich in functional assessment. This result, if validated by larger scale studies not only in the area of listing criteria but also in the vocabulary used in, for example, assessment instruments, raises questions about the roles ICF and SNOMED CT should play in standardizing terminologies used in functional assessment. In an ideal world, ICF and SNOMED CT are harmonized so that SNOMED CT’s more granular vocabulary can enrich and complement ICF’s systematic descriptions of functions, disabilities, and factors that affect them.
The results reported in this paper should be considered preliminary, as the sample size of criteria used in the experiment is small and the concept decomposition, terminology mappings, and characterization of criteria as functional and non-functional had not been checked by experts on disability determination and terminologies. Nevertheless the qualitative results probably hold. SNOMED CT is a much larger terminology than ICF. The range and granularity of SNOMED CT lead to higher coverage rate of concepts used to define Mental Disorders criteria.

The size and complexity of SNOMED CT, however, means that finding the right mappings is a more error-prone process. Furthermore, SNOMED CT allows multiple decomposition and mapping possibilities. In particular, concepts described by SNOMED CT clinical findings often can be defined by a combination of observable entity and qualifier values. For example, the Mental Disorders concept "Impaired memory" can be mapped to either “Impaired memory” (clinical finding 386807006) or a Change function combination of “Impaired” (qualifier value 260379002) and “Memory function” (observable entity 303116000). A more granular mapping (e.g., using the observational entity and qualifier value combination instead of a single clinical finding concept) will drive up the coverage ratio.

The innovation of this work involves the development of a structured decomposition method for identifying appropriate concepts, qualifiers, and qualifier values in complex criteria that should be mapped to standard terminologies. This decomposition defines the semantic roles that the entities play in helping to define the meanings of these concepts. This decomposition is essentially a kind of post-coordination. As such, it can be formalized in description logic. For example, a qualified concept consisting of a root concept C, a qualifier Q, and a qualifier value V is essentially an OWL class expression (using the Manchester Syntax: C AND (Q some V)). Given this interpretation of decomposition, we can generate logic expressions that capture much of the meanings of the disability criteria, as was described in [4]. Given appropriate data, we can use a description-logic reasoner to classify the data to determine whether they satisfy the logical expressions, potentially providing a method to help determining whether a claimant is eligible for disability benefits.

Conclusions

We developed a structured decomposition method for identifying appropriate concepts, qualifiers, and qualifier values that occur in eligibility criteria and that should be mapped to standard terminologies. A proof-of-concept application of this method to the Adult Mental Disorders chapter of SSA’s Blue Book listing criteria shows that SNOMED CT has significantly higher coverage of functional and non-functional concepts than ICF. While the method is applicable to terminology coverage studies involving complex criteria, the specific findings of the proof-of-concept experiment need to be further validated by disability determination and terminology experts in larger studies.

Acknowledgment

This work was supported by Interagency Personnel Agreements funded by the Social Security Administration (SSA). We gratefully acknowledge the assistance provided by Derek Wang, Rosemary Hall, Dr. Laurence Desi Sr, Bob Hastings, Joe Herendeen, Ronald Lee, Michele Schaefer, Cheyl Williams, and Jim Twist from SSA and John Hough from the Center for Disease Control and Prevention.

References


Use of Patient Portals for Personal Health Information Management: The Older Adult Perspective

Anne M. Turner, MD, MLIS, MPH1, Katie Osterhage, MMS1, Andrea Hartzler, PhD2, Jonathan Joe, BS1, Lorelei Lin, BA1, Natasha Kanagat, MPH1, George Demiris, PhD1
1University of Washington, Seattle, WA; 2Group Health Research Institute, Seattle, WA

Abstract
The personal health information management (PHIM) practices and needs of older adults are poorly understood. We describe initial results from the UW SOARING project (Studying Older Adults & Researching Information Needs and Goals), a participatory design investigation of PHIM in older adults (60 years and older). We conducted in-depth interviews with older adults (n=74) living in a variety of residential settings about their management of personal health information. A surprising 20% of participants report using patient portals and another 16% reported prior use or anticipated use of portals in the future. Participants cite ease of access to health information and direct communication with providers as valuable portal features. Barriers to the use of patient portals include a general lack of computer proficiency, high internet costs and security concerns. Design features based on consideration of needs and practices of older adults will facilitate appeal and maximize usability; both are elements critical to adoption of tools such as patient portals that can support older adults and PHIM.

Keywords:
Patient portals; older adults; health information management

Introduction and Background
A patient portal is a type of personal health record (PHR) that is connected to an electronic health record (EHR) system. Patient portals provide a secure website through which patients can access their clinical data. They are a key component of most EHR architectures and an important focus of meaningful use because of their potential to streamline the delivery of patient-centered health care. Features of patient portals may include secure messaging, after-visit summaries, medication lists, allergy lists, laboratory results, and appointment scheduling. When used effectively, patient portals can empower consumers by enabling active management of their own care. However, we know little about how patient portal use fits into the broader personal health information management (PHIM) practices of various groups, such as older adults.

Within the general U.S. population, there is considerable interest in patient portals. Nearly 80% of consumers participating in a 2008 Deloitte Survey of Health Consumers expressed interest in accessing an integrated medical record that provides information on personal test results, doctor visits, and hospital stays. Of those participants, 75% expressed a desire for online portals that facilitate appointment schedules, email communication with physicians, access to test results and medical records. Other research also shows similar findings with regard to interest in using online personal health records. However research about the usage patterns of various users, as well as the association between use of patient portals and quality of care and health outcomes, is sparse.
In particular, research evidence about PHRs and older adults is limited. Witry et al. examined views of family practice physicians and staff about the benefits, barriers, and use of PHRs by older adults, in relation to medication use. They concluded that the family practice physicians did not have a complete understanding of the benefits that PHRs can offer patients. Several studies highlight the challenges that older adults face when using PHRs, such as physical or cognitive limitations and low computer literacy. Providers have also raised concern that use of PHRs could introduce privacy risks to patients. Other providers worry that older adults may be especially vulnerable to “getting scammed” while using electronic or online PHRs.

These findings demonstrate providers’ hesitancy to promote the use of PHR among older patients, whom they see as susceptible to privacy violations. Security has been examined by Weitzman and colleagues in the context of personally controlled PHRs. By studying a PHR called Indivo, the researchers identified important concerns of older adults about a possible breach in security. Older participants felt they had “less to lose” than younger participants, if a security breach occurred. However, the same participants expressed worry that information disclosure through a PHR could impose an emotional burden on their family members.

Other researchers have focused on older adults’ use of PHRs for a range of functions, such as home medication management and information sharing with providers. For example, Kim et al. explored the use and utility of PHRs in a low-income population of older adults. The 33-month study involved 44 older adult residents of a federally funded housing facility. Researchers assessed use of and user satisfaction with a web-based PHR, the Personal Health Information Management System (PHIMS). Use among the residents was low, with only 13% of eligible residents using the system and, of those, about half used the system on only one occasion. This study concluded that the majority of the low-income elderly would not benefit from PHRs, due to poor technical skills, low health literacy and limited physical/cognitive abilities. Thus, older adults face considerable barriers in their adoption and use of PHRs and other information technology that could support PHIM. In addition, patient portal technologies have not been designed with consideration to older adults’ needs and preferences, which may explain why older adults have been slower to adopt these new technologies for managing personal health information.

Our SOARING project, a 5-year Agency for Healthcare Research and Quality (AHRQ) funded investigation, is designed to address this gap by establishing an understanding of the PHIM practices and needs of older adults. Based on Grounded Theory and the ecological framework of the Balance Theory, we sought to identify current health information practices and needs among older adults living in a variety of residential settings. Using focus groups, in-depth interviews with longitudinal follow-up, and participatory design with older adults and their key stakeholders, our goals are to develop older adult-centered guidelines to assist developers in the design of useful and usable health information management tools that better serve older adults.

We report here on our initial findings from semi-structured interviews concerning older adults’ use of patient portals for accessing and organizing health information.

**Methods**

**Recruitment**

As part of our larger SOARING study (http://www.nwphp.org/research/projects/current/health-information-management-older-adults), we conducted semi-structured interviews with older adults (≥ 60 years) to better understand how older adults manage their health information. We recruited participants from residential communities, assisted living facilities, and independent residences in King County, Washington. Recruitment efforts included presentations, flyers and key contacts at senior centers, community organizations, and older adult communities. We employed purposive sampling to ensure a diverse representation of age, gender, socio-economic status (SES), racial and ethnic backgrounds, across a range of living situations. With regard to diversity of living situation, for instance, we recruited several participants who were self-described as either currently homeless or recently homeless (n=3).
Inclusion criteria included age 60 years and older, the ability to read and speak English, and lack of cognitive impairment, measured by a score of 4 or higher on the 6-item screener to identify cognitive impairment. The University of Washington IRB approved all study procedures (IRB #45853).

Semi-structured interviews

We conducted semi-structured, 60-90 minute interviews with each participant. In most instances, interviews took place at a participant’s place of residence. After obtaining consent, we audio recorded the interview, photographed artifacts associated with PHIM, and made field notes. The interview session consisted of a demographic survey followed by a series of semi-structured questions concerning health and the use and organization of personal health information. Topics covered in the interview guide included: health conditions, management of health conditions, interactions with healthcare providers and corresponding materials received, health-related record keeping, health information seeking, and use of patient portals. Examples of specific questions include: Do you keep any records related to your health? What information (if any) do you keep track of to keep yourself healthy? What tools do you use to track your health information?

Initially, we asked participants to describe their general experience with technology and we did not focus on any specific technology or tools. However, over time, due to the frequency with which participants mentioned patient portals, we added structured and semi-structured questions regarding patient portals. In introducing this section of the interview, we described patient portals to participants as, “Websites used to access your health records or to communicate with your provider team.” Local examples of Epic MyChart and Cerner PHR were shared with participants. We probed with questions to explore barriers and facilitators to patient portal use among participants.

Analyses

Interview audio recordings were transcribed by a professional transcription service. Two research team members reviewed the transcripts and notes for discussions related to portal use. Demographic survey data were summarized with descriptive statistics using Microsoft Excel.

Results

Participants

Demographics: We interviewed 74 participants. Of those interviewed, 20% (15/74) reported using a patient portal, such as Epic MyChart™ (Epic Systems Corp, Verona, WI), to help manage their health information. Table 1 compares demographics of the 15 portal users with demographics of the remaining 59 participants who report not using a patient portal (“portal nonusers”). Portal users ranged in age from 61 to 93 years, and most lived independently in a private residence (60%) and had college education or higher (67%). Although portal nonusers were similar in age, fewer were college educated (53%) and more lived in retirement or assisted living facilities (74%). Whereas 28% of portal nonusers reported having an informal or unpaid caregiver, only 7% of portal users reported this. Eighty percent of the portal users identified themselves as white.
Table 1. Participant demographics and living situation

<table>
<thead>
<tr>
<th></th>
<th>Portal Users (n=15)</th>
<th>Portal Nonusers (n=59)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (SD)</td>
<td>76 (7.4)</td>
<td>78 (10.2)</td>
</tr>
<tr>
<td>Female</td>
<td>87%</td>
<td>59%</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some high school or less</td>
<td>0%</td>
<td>5.1%</td>
</tr>
<tr>
<td>High school graduate</td>
<td>6.7%</td>
<td>13.6%</td>
</tr>
<tr>
<td>Some college</td>
<td>26.7%</td>
<td>28.8%</td>
</tr>
<tr>
<td>College graduate</td>
<td>20.0%</td>
<td>33.9%</td>
</tr>
<tr>
<td>Post graduate</td>
<td>46.7%</td>
<td>18.6%</td>
</tr>
<tr>
<td>Living situation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retirement community</td>
<td>20%</td>
<td>49%</td>
</tr>
<tr>
<td>Assisted living facility</td>
<td>13%</td>
<td>25%</td>
</tr>
<tr>
<td>Private residence</td>
<td>60%</td>
<td>17%</td>
</tr>
<tr>
<td>Other (e.g., homeless)</td>
<td>7%</td>
<td>9%</td>
</tr>
</tbody>
</table>

Note: Later in this section, we discuss a group of “potential portal users.” In this table, portal nonusers (n=59) include these twelve participants we later mention as potential portal users.

Technology use and experience: Our preliminary analysis found that most portal users (93%) reported using a computer 6-7 days per week and 47% rated themselves as “very experienced” computer users. All reported having learned to use a computer 10 or more years ago and having Internet access where they live. In contrast, the majority of portal nonusers reported having used computers less than 6 days per week (30%) or not at all (36%), 25% lacked Internet access where they lived, and 19% reported they had not learned to use a computer. Among the 52 participants (portal users and nonusers) who use computers, the most common use of computers was emailing and browsing the Internet.

From the interviews, two themes emerged with regard to patient portals: ‘facilitators,’ characteristics, uses or other attributes that encouraged or eased utilization of a patient portal, and ‘barriers,’ individual or systems level obstacles to patient portal use.

**Facilitators of portal use**

The majority of patient portal users made positive comments about the specific portal they used. Easy access to health information, more direct communication with providers, and the ability to make appointments online were cited as valued features.

Many (9 of 15) portal users mentioned ease of access to health information, as illustrated in the quotes below:

- *...they have the whole record there. That’s what I love about it.*
- *The doctor has email, and you can get the lab results from there. So anything I need to know, like, if I think my potassium is too low or something, I can check it and see.*

Direct communication with providers and the ability to easily make an appointment were cited as other positive features of patient portals:

- *You can email back and forth with your doctor if you don’t necessarily need an appointment.*
- *I like the fact that I could, rather than calling and running the risk of a note being taken wrong or misunderstanding what I said, it’s there. It’s in my own typing. It’s there. And I can send my doctor notes about my blood pressure, or ask her a quick question. I love computer medical technology.*
- *...it’s a faster and better way to communicate with my doctor about things. You see, before, the only way I could communicate with my doctor was call over to adult medicine and leave a message if I had to.*

From portal users, we heard many positive comments regarding the ease of use, such as:

- *...it’s been really good, and helpful, and easy to utilize. They made the system user friendly.*
I like that it is easy to access and easy to navigate.

*MyChart is really cool. ... I can make* an appointment online with my doctor. All my doctors there are listed. I just click on the one I want.

Regarding management of health information, several participants mentioned a dramatic reduction in their health information record keeping with portal use.

- I used to file things away in file folders, but now I pretty much just go check, and if I want to find out what my blood pressure’s been for the last five years, all I have to do is go online and they have these nice little lists.... So they made me kind of lazy really, as far as keeping other records. Because it’s right there.

**Barriers to portal use**

Potential portal users: Of the 59 nonusers, 20% (16% of the total participants interviewed) mentioned some relationship to patient portals. Four participants stated they had previously used a portal, three participants had use of a portal through a family member, and two participants indicated they would like to use a patient portal in the future.

Within this potential portal user group, problems with logging in were provided as reasons for no longer using patient portals:

- *Because I don’t use it all the time, I forget what the password is and what the user I.D. is.*

- I gotta make sure I still have my password and everything so I can get back on it.

Some of the potential portal users expressed positive attitudes towards patient portals, but cited the cost of maintaining internet access as a barrier to their use:

- The experience was great. If I had internet at home, if I could afford internet, I would definitely use MyChart on a regular basis.

- *But for many people, we cannot afford the internet. I can’t afford $29 a month.*

Portal Nonusers: Many of the remaining participants classified as portal non users had never heard of a patient portal and/or expressed negative attitudes towards the use of portals. Participants who did not use patient portals primarily cited personal reasons, such as general aversion to or difficulty using computers, and security concerns:

- My brain will only stand so much confuzzling.

- *I’m not interested [in patient portals]. I don’t like any of my things on the web or whatever you call it. I’m against all that.*

- *Well, I’m paranoid of computers. People get their identity, they find out their weaknesses, and they swoop in, and I really don’t feel like broadcasting all my stuff on the internet.*

Of note, factors intrinsic to specific patient portals, such as difficulty navigating a site or interface design, were not mentioned by any participants.
Discussion

Over the last twenty years, there has been growing emphasis on the involvement of health consumers in their own healthcare. Important to the success of the consumer health movement is accurate, accessible, and understandable health information to assist with treatment and health decisions. Older adults are the largest consumers of health care and expend the greatest proportion of US health care dollars. Thus, their participation is key to the success of this consumer movement towards greater health autonomy.

Despite reported barriers to the use of PHR technologies by others\textsuperscript{16, 17}, 20% of the older adults we interviewed use patient portals to manage their personal health information. These participants express enthusiasm about the access to health information, direct communication with providers, and ease of scheduling that patient portals offer. These factors appear to serve as important facilitators for adoption of patient portals by older adults.

Our finding that those older adults who use patient portals do so because the use of the portal facilitates efficient and easy access to health information, medical reports, and clinical appointments, is consistent with prior studies investigating older adult adoption of new technologies, such as automated teller cards and cell phones\textsuperscript{18, 19}. These studies indicate that older adults generally lag behind younger adults in adopting new technologies\textsuperscript{16, 17}, but will adopt new technologies when they feel these technologies enhance convenience, are useful to their daily lives, and are easy to use\textsuperscript{21, 22}. An important insight our study adds to prior work is that patient portals appear to play a small but substantial role in facilitating PHIM in older adults.

As in prior research, we identified several barriers to the use of patient portals by older adults. Our results indicate that these barriers are not only limited to the individual end user level, such as computer knowledge or prior training, but also exist at a systemic level, including infrastructure that hinders Internet access and lack of resources within residential facilities to equip residents with tools that facilitate access.

Despite the promising insights our findings offer, generalizability is limited due to the small number of older adults living in King County, Washington. Although a sizeable proportion of participants we engaged used patient portals, this proportion and its composition could differ in other geographic locations. Portal users in our study were highly educated and lived independently, which complements groups of older adults in prior studies\textsuperscript{9}. It is not clear the extent to which portal use may result from or promote independent living. Expanded studies are needed to determine the general penetration of patient portals in this population, factors that contribute to portal use by older adults, and associations between patient portal use and independent living.

Our findings provide initial insights into the PHIM needs and practices of older adults regarding patient portals. Further research is needed to better understand what factors related to form (mobile phones, tablets, desktop computers) and function are most important to include in patient portals when designing user interfaces that facilitate management of personal health information by older adults. A better understanding of the role of caregivers in portal use will be the focus of future studies involving interviews we will conduct with key stakeholders. By establishing an understanding of the PHIM practices and needs of older adults, we can substantially inform the design of supportive technologies and promote the adoption of patient portals by this important population.

Conclusion

Despite broader issues of accessibility, such as computer use and Internet access, study findings highlight the small but substantial role of patient portals as a platform to facilitate management of personal health information among older adults. The use of patient portals demonstrates their growing potential and role in helping older adults maintain...
the health, wellness, and independence they desire. Design features based on consideration of needs and practices of older adults will facilitate appeal and maximize usability, which are critical to adoption.

References


Acknowledgments

This research was funded by the Agency for Healthcare Research and Quality (AHRQ) grant # R01 HS022106. The content is solely the responsibility of the authors and does not necessarily represent the official views of the AHRQ. We would like to thank the older adults who participated in this study, as well as the residential agencies and organizations that assisted in recruitment. In addition, the authors thank Julie Loughran and Beryl Schulman for their help in preparing this manuscript.

Address for correspondence

Contact Dr. Anne Turner at amturner@uw.edu
The State and Trends of Barcode, RFID, Biometric and Pharmacy Automation Technologies in US Hospitals

Raymonde Charles Y. Uy, MD, MBA, Fabricio P. Kury, MD, Paul A. Fontelo, MD, MPH
National Library of Medicine, Bethesda, MD 20894

Abstract

The standard of safe medication practice requires strict observance of the five rights of medication administration: the right patient, drug, time, dose, and route. Despite adherence to these guidelines, medication errors remain a public health concern that has generated health policies and hospital processes that leverage automation and computerization to reduce these errors. Bar code, RFID, biometrics and pharmacy automation technologies have been demonstrated in literature to decrease the incidence of medication errors by minimizing human factors involved in the process. Despite evidence suggesting the effectiveness of these technologies, adoption rates and trends vary across hospital systems. The objective of study is to examine the state and adoption trends of automatic identification and data capture (AIDC) methods and pharmacy automation technologies in U.S. hospitals. A retrospective descriptive analysis of survey data from the HIMSS Analytics® Database was done, demonstrating an optimistic growth in the adoption of these patient safety solutions.

Introduction

In 1999, the Institute of Medicine estimated that the hospital mortality rate due to medication errors is as much as 98,000 patients per year, thereby making these errors a major public health concern. In their report, the committee recognized that addressing patient safety is the key component for the delivery of quality healthcare. It recommended improvements that have to be made in hospital systems and processes to reduce injuries. Different safety systems support the five rights of medication administration (right patient, right drug, right time, right dose, and right route) at various steps of the medication administration process. While electronic health record systems and computerized physician order entry are primarily focused on preventing order errors in the prescribing, transcribing and documentation steps, additional errors can transpire in the dispensing, and administering phases. It is estimated that the majority of medication errors occurs at the prescribing (49%) and administration (26%) steps.

The years following the IOM report exhibited some small but insufficient progress in addressing medication errors through changes in health systems and policies. In 2004, the U.S. Food and Drug Administration (FDA) initiated a ruling to require all human medications and biological product labels to contain barcodes to help prevent medication errors and avert costs related to these adverse events. In 2012, as additional support to the FDA ruling, the Centers for Medicare & Medicaid Services (CMS) required hospitals to begin tracking medications starting from when a medication order is initiated, until its administration to the patient. The CMS measures suggest the implementation of assistive technologies such as automatic identification and data capture (AIDC) methods like barcoding, radio frequency identification (RFID) and biometrics in conjunction with pharmacy automation technologies like automated dispensing machines, carousels, and robotics, as part of Stage 2 Meaningful Use Core Measures.

Various AIDC methods have been demonstrated in literature to be effective in the reduction of medication errors and other adverse events. Barcode technology is used in various departments and processes in the hospital such as the laboratory, pharmacy, radiology, and medication administration. The technology typically uses a handheld barcode reader that registers and documents events of contact between medication, equipment, and healthcare personnel. Bar code medication administration (BCMA) is an integral part of preventing medication errors by making sure that the right patient is receiving the right medication at the point of administration. It is reported that BCMA systems have reduced the incidence of medication errors by more than 50%, and the risk of adverse drug events by 11% or approximately 20 events per day. The use of BCMA is also required in a closed-loop medication administration environment, which is the primary criteria for Stage 5 of the Healthcare Information and Management Systems Society (HIMSS) EMR Adoption Model (EMRAM), in addition to various levels of EMR system capabilities. It requires an electronic medication administration record (eMAR) that uses an AIDC method, which is integrated with the...
computerized provider order entry (CPOE) and pharmacy information systems. A closed-loop environment significantly reduces medication errors, and increases patient identity confirmation before medication administration17.

Bar code systems have also been applied to hospital laboratory processes. In neonatal intensive care units (NICU), barcoding in breast milk administration has been used to ensure that fragile, hospitalized infants who are separated from their mothers receive the right expressed milk at the right time from the right parent, eliminating the risk of transmitting infectious diseases caused by human immunodeficiency virus, cytomegalovirus, and hepatitis B virus16-21. Barcoding has also been used to reduce adverse events and errors from the transfusion of blood products, which is estimated to occur once every 12,000 units transfused in the United States. Over 50% of the mortality from transfusion-related injuries is attributed to errors in patient verification22-24. Implementation of barcoding for transfusion verification has been demonstrated to be effective in preventing mismatched transfusions that may lead to the transmission of HIV, hepatitis B/C, and severe reactions such as acute hemolysis from ABO incompatibility24-26.

RFID technology has also been used in hospitals for the same purposes as barcodes. The technology uses radio waves for collecting and transferring patient data25. Some of its advantages over barcode technology include the elimination of the “line-of-sight” requirements of barcode scanners, and the capability to program RFID devices. RFID has been demonstrated to be effective in supporting patient safety, eliminating medication errors, and other adverse events related to patient misidentification28-30.

Other newer technologies enhance patient safety by providing patient security through identity verification systems that use biometrics. Biometrics are measurable characteristics of human beings that are unique to each individual. Biometric devices and their accompanying software in healthcare institutions permit the automatic authentication of patient and provider identity for different purposes such as secure EHR system access, and patient verification31. The most common hospital implementation of biometrics are the use of fingerprint and iris scanning32, 33. The unique authentication methods of biometrics make it difficult to mismatch and forge identities since no two irises or fingerprints are the same.

In the pharmacy, different hardware systems that work in conjunction with hospital information systems play a major role in assuring that the right drug, the right dosage and instructions are prepared for the right patient. Automated dispensing machines (ADM), or automated dispensing cabinets are the most commonly used devices for decentralized medication dispensing34. These machines are storage devices that automate and track medication distribution at the point-of-care. ADMs have been shown in the literature to have a moderate effect in reducing medication errors by automating the dispensing process34, 35. Carousels on the other hand are centralized medication storage and retrieval systems designed as a series of revolving shelves set on rails has been seen to reduce filling or dispensing errors by automating medication dispensing in the pharmacy36. Lastly, stationary robotic compounding and dispensing systems that work in conjunction with ADMs and carousels to further increase the accuracy of dispensing the correct drugs, dosage and quantities to the right patients, have demonstrated clear benefits in patient medication safety37. These pharmacy automation technologies all use a form of AIDC method to verify that medication orders are correctly linked to patient records, and monitor inventory supplies.

Motivated by the important role and potential that AIDC methods and pharmacy automation implementations possess in reducing medication errors, we aimed to determine the state and trends of adoption of these technologies in U.S. hospitals using a retrospective descriptive analysis of survey data from the HIMSS Analytics® Database
Methods

Data Source

Data from the Healthcare Information and Management Systems Society (HIMSS) Analytics® Database was used in this study. The HIMSS Analytics® Database contains survey data on the use, implementation and planning status of health IT hardware, software and infrastructure of more than 5,400 non-federal U.S. hospitals, which is included in its catalogue of nearly 40,000 U.S. healthcare facilities. The database is noted to be the “most comprehensive database” for hospital IT adoption, representing nearly all non-federal hospitals with greater than 100 beds, and more than 90% of all U.S. hospitals. The annually updated database originally started as the Dorenfest 3000+ Database in 1986, and was integrated in 1998 with data from U.S. Integrated Health Delivery Systems, to form the Dorenfest IHDS+ Database. It was then acquired by HIMSS Analytics, a non-profit subsidiary of HIMSS, in 2004 to become the HIMSS Analytics® Database. The latest version of the database used in this study contains data for 2012, which was made available at no charge to academic researchers in July, 2014.

Measurement of Technology Adoption

The HIMSS Analytics® Database was used to explore specific data elements pertaining to AIDC and pharmacy automation technology users. In each data element, survey respondents indicated whether the technology is currently being used, with some elements containing information on plans for future adoption. The following hierarchy tree loosely represents the availability and organization of the database relating to the topic of interest. The HIMSS survey definition of each data element can be found in Appendix A.

I. Barcoding
   A. Laboratory Department
      • Breast Milk Administration
      • Transfusion Verification
   B. Pharmacy Department
   C. Radiology Department
   D. Medication Administration
      • Complete Closed-Loop Medication Administration

II. Radio-frequency Identification (RFID)
   A. Laboratory Department
   B. Pharmacy Department
   C. Radiology Department
   D. Medication Administration

III. Medication Administration Processes

IV. Biometrics
   A. Fingerprint Scanning
   B. Iris Scanning

V. Pharmacy Automation
   A. Automated Dispensing Machines (ADMs)
   B. Carousels
   C. Robotics

Data Analysis

The Microsoft Access files of the HIMSS Analytics® Database, from 2008 to 2012, were queried using SQL from inside an R script. The R script gathered the results of the SQL queries, assembled together the data from separate years, and exported Office Open XML spreadsheets (XLSX files). All source code used in this paper is available for free reuse, commentary and contribution at http://github.com/fabkury/itsos. Exploratory analyses were done to inspect the technology adoption rates in U.S. hospitals. Hospitals with missing values were not considered in the analysis of each data element individually.
Results

Barcoding Use

Bar code medication administration (BCMA) had the highest growth in adoption, averaging an increase of about 7% per year from 2008 to 2012 compared to 6.4%, 2.2%, and 1.15% for Pharmacy, Laboratory, and Radiology departments respectively (Figure 1). Barcoding use in the Laboratory department had the highest hospital adoption at 84.2% in 2012, compared to 73.9%, 58.1%, and 50.8% for Pharmacy, medication administration, and Radiology respectively. Within hospitals that use barcoding in the Laboratory (n=4509), 7% (n=324) use the technology for transfusion verification, while only 2% (n=89) use it in breast milk administration in 2012. In the same year, within hospitals that use BCMA (n=3114), 67.8% (n=2110) attest to having a complete closed-loop medication administration system, while 12.3% (n=384) indicate that they do not, and 19.9% failed to provide the detail. Table 1 includes a summary of adoption rates for bar coding, and Figure 1 for annual trends.

RFID Use

RFID medication administration had the highest growth in adoption, averaging an increase of about 0.4% per year from 2008 to 2012 compared to 0.25%, 0.24%, and 0.14 for Laboratory, Pharmacy and Radiology departments respectively (Figure 2). Similarly, RFID medication administration had the highest adoption rate in 2012 at 1.87% compared to 1.57%, 1.55%, and 1.12% for Laboratory, Pharmacy and Radiology departments respectively. A summary of the adoption rates is shown in Table 1.

Table 1. Percent use (%) of Bar code and RFID technologies by Department

<table>
<thead>
<tr>
<th></th>
<th>Year</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcode</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Laboratory</td>
<td>75.3</td>
<td>79</td>
<td>81.3</td>
<td>82.2</td>
<td>84.2</td>
<td></td>
</tr>
<tr>
<td>2. Pharmacy</td>
<td>48.3</td>
<td>55.7</td>
<td>62.1</td>
<td>67</td>
<td>73.9</td>
<td></td>
</tr>
<tr>
<td>3. Medication Administration</td>
<td>29.9</td>
<td>41.3</td>
<td>45.1</td>
<td>50.3</td>
<td>58.1</td>
<td></td>
</tr>
<tr>
<td>4. Radiology</td>
<td>46.2</td>
<td>47.9</td>
<td>50</td>
<td>50</td>
<td>50.8</td>
<td></td>
</tr>
<tr>
<td>RFID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Medication Administration</td>
<td>0.27</td>
<td>0.7</td>
<td>1.14</td>
<td>1.37</td>
<td>1.87</td>
<td></td>
</tr>
<tr>
<td>2. Laboratory</td>
<td>0.57</td>
<td>0.69</td>
<td>0.91</td>
<td>1.09</td>
<td>1.57</td>
<td></td>
</tr>
<tr>
<td>3. Pharmacy</td>
<td>0.49</td>
<td>0.59</td>
<td>0.83</td>
<td>1.05</td>
<td>1.55</td>
<td></td>
</tr>
<tr>
<td>4. Radiology</td>
<td>0.54</td>
<td>0.61</td>
<td>0.69</td>
<td>0.78</td>
<td>1.12</td>
<td></td>
</tr>
</tbody>
</table>
Medication Administration Process

For all hospitals in 2012 that either use bar code or RFID in their medication administration process (n=2901), 97.1% (n=2818) of the hospitals have tags on the medications, 89.4% (n=2594) have them on their patients, and 59.2% (n=1718) on nurses. Patient tagging had the highest growth in adoption per year at 3.4%, followed by nurse (3.38%) and medication (0.96%) tagging and from 2008 – 2012 (Figure 3).

Biometrics

The use of fingerprint scanning had an average annual adoption increase of 1.23% per year from 2008, leading to a total of 15.9% (n=871) adoption within all hospital respondents (n=5467) in 2012. Iris scanning technology has an average annual adoption rate of 0.02% from the same time frame, and is only currently being used in 13 hospitals (0.02%) in 2012. Only 2.49% (n=136) hospitals in 2012 plan to expand or adopt the fingerprint technology in the following years, in contrast to only 12 hospitals (0.22%) for iris scanning.

Pharmacy Automation

The adoption of automated dispensing machines increased at an average annual adoption rate of 2.08% from 2008, leading up to an 81% hospital use in 2012. ADMs are predominantly used in the Medical or Surgical departments, followed by the Emergency department and the operating rooms (Figure 4). In contrast, robotic technology average annual adoption rate is around 0.04% within the same 4-year timeframe, leading to 7.88% hospital use in 2012. Carousel hospital use increased from 2.9% in 2010 to 5.67% in 2012. About 20.32% of hospitals that are already using carousels plan to expand and acquire more.

Table 2. Percent use (%) of Biometrics and Pharmacy Automation

<table>
<thead>
<tr>
<th></th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biometrics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Fingerprint Scanning</td>
<td>11</td>
<td>12.8</td>
<td>14.3</td>
<td>15.1</td>
<td>15.9</td>
</tr>
<tr>
<td>2. Iris Scanning</td>
<td>0.15</td>
<td>0.23</td>
<td>0.23</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td><strong>Pharmacy Automation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Automated Dispensing Machines</td>
<td>72.7</td>
<td>75.9</td>
<td>78.9</td>
<td>80.4</td>
<td>81</td>
</tr>
<tr>
<td>2. Carrousels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Robotics</td>
<td>7.72</td>
<td>7.98</td>
<td>8.2</td>
<td>8.09</td>
<td>7.88</td>
</tr>
</tbody>
</table>
Discussion

Overall, we found that from 2008 to 2012, medication administration had the highest annual growth rate for both bar code (7%) and RFID (0.4%) technologies. This may be due to healthcare legislation such as the HITECH Act in 2009, which stimulated investments in health IT, and set deadlines to meet specific meaningful use criteria. The data also shows that about 67.8% of hospitals with BCMA have a closed-loop system. This raises the optimism for more hospitals to reach Stage 5 of the HIMSS EMRAM.

Bar code adoption in laboratory (84.2%) and pharmacy (73.9%) departments in 2012 are high, which may be attributed to regulations by the FDA, and the amount of inventory tracking inherent in both departments. We also notice the low adoption rates for breast milk administration and transfusion verification, which may be attributed to the relatively low error rates compared to general medication administration, and the lack of consensus and economic resource. Adoption of RFID technology is also generally low, which is consistent with previous studies explaining costs and negative perceptions as major barriers of implementation.

Details on the utilization of bar code and RFID tags on patients, medications, and nurses in 2012 show both high rates for medications (97.1%) and patient (89.4%) tags, but moderate use for nurses (59.2%). This may have implications on the requirements of closed-loop medication administration and information in the eMAR because a record containing the identity of the healthcare provider who dispenses and administers the medication is important in tracking and record transparency.

For biometrics use in hospitals, the data shows a slower annual adoption of both fingerprint and iris scanning technologies (1.23% and 0.02% respectively), with comparatively low overall adoption. Low adoption may be due to the costs in implementing biometrics within existing EHR systems and workflows. Pharmacy automation technologies such as ADMs seems to show a steady growth (2.8%), with high hospital adoption (81%) in 2012. In contrast, adoption is relatively low for carousels and robotics (5.67% and 7.88% respectively), which is most likely due to the hardware costs.

In summary, although the benefits of AIDC methods and pharmacy automation technologies in reducing medication errors seem to be increasing, current adoption trends, careful consideration of individual hospital systems, costs, and clinical workflow should guide administrative decisions leading to greater adoption. Organizational leadership and the cooperation of hospital staff will continue to be important in the ongoing adoption of these technologies. Findings from this study may provide decision makers with a benchmark for strategic planning and deployment of these technologies for raising the quality of healthcare through the improvement of patient medication safety.
Limitations
The main limitations of the HIMSS Analytics® Database are similar to other studies based on data from surveys. The data relies on accurate self-reporting from hospital administrators via completion of a phone interview and an IT inventory survey. Survey respondents who represent their healthcare institution may not have the information to answer some specific parts of survey questions, or may decline to volunteer the information, leading to varying response rates for each data element. Some data elements lack more granularity and other specific process details that may open new opportunities for research and analyses. In common with many surveys, the time period during data collection may not accurately reflect the current state of affairs in these organizations.

Regardless of its limitations, to our knowledge, the HIMSS Analytics® Database is currently the most complete survey of health IT adoption and implementation in nearly all non-federal U.S. hospitals. It is a valuable resource that continues to be refined and expanded every year, providing researchers with trends that contribute to more strategic hospital investments and planning.

Conclusion
The increasing adoption of AIDC methods and pharmacy automation technologies across all U.S. hospitals sizes demonstrates interest in ensuring patient medication safety towards the improvement of quality of care. A comprehensive knowledge of the adoption rates, trends and evidence demonstrating the effects of implementing these technologies may contribute to individual hospitals’ strategic decision-making. The database provided by HIMSS Analytics™ with the support of the Dorenfest Institute are capable of providing analytic data on these and other matters that may serve as indicators for healthcare access and delivery. The same knowledge may stimulate the development of future health policies supporting the five rights of medication administration.

Acknowledgement
We gratefully acknowledge HIMSS Analytics™ for access to the data, especially Jennifer Horowitz (Senior Director of Research) of HIMSS North America for her valuable assistance with the database. This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and Lister Hill National Center for Biomedical Communications (LHNCBC). This research was also supported in part by an appointment to the NLM Research Participation Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy (DoE) and the NLM.

Disclaimer
The views and opinions of the authors expressed herein do not necessarily state or reflect those of the National Library of Medicine, National Institutes of health or the US Department of Health and Human Services.

Competing Interests
The authors declare no competing interests.
References

**APPENDIX A. HIMSS Data Element Definitions**

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Barcoding</strong></td>
<td><strong>Laboratory</strong> Bar code technology used in the laboratory department to improve the efficiency of operations for functions such as specimen identification, specimen collection, and specimen processing.</td>
</tr>
<tr>
<td></td>
<td><strong>Pharmacy</strong> Bar coding is used by the pharmacy department for inventory control of drugs.</td>
</tr>
<tr>
<td></td>
<td><strong>Radiology</strong> Bar code technology used in the radiology/imaging department(s) to improve the efficiency of operations of functions such as patient tracking, film tracking, and the completion of imaging services.</td>
</tr>
<tr>
<td></td>
<td><strong>Medication Administration</strong> Barcode technology used by nursing services to improve the efficiency of operations such as patient identification, nurse identification, medication identification, and closed loop medication administration processes that improve patient safety.</td>
</tr>
<tr>
<td></td>
<td><strong>Closed-Loop Medication Administration</strong> An environment where the medication process is electronic from initial entry by physicians using CPOE, to pharmacies for order validation and bar coding the medications, to the automatic dispensing machines, to the actual administration of the medication at point of care by the nurse where the nurse scans patient bar code and the medication bar code which initiates clinical decision support for the five rights of medication administration.</td>
</tr>
<tr>
<td><strong>RFID</strong></td>
<td><strong>Laboratory</strong> RFID technology used in the laboratory department to improve the efficiency of operations for functions such as specimen identification, specimen collection, and specimen processing.</td>
</tr>
<tr>
<td></td>
<td><strong>Pharmacy</strong> RFID is used by the pharmacy department for inventory control of drugs.</td>
</tr>
<tr>
<td></td>
<td><strong>Radiology</strong> RFID technology used in the radiology/imaging department(s) to improve the efficiency of operations of functions such as patient tracking, film tracking, and the completion of imaging services.</td>
</tr>
<tr>
<td></td>
<td><strong>Medication Administration</strong> RFID technology used by nursing services to improve the efficiency of operations such as patient identification, nurse identification, medication identification, and closed loop medication administration processes that improve patient safety.</td>
</tr>
<tr>
<td></td>
<td><strong>Medication Administration Processes</strong> Elements that are bar coded or have an RFID tag in the medication administration process.</td>
</tr>
<tr>
<td><strong>Biometrics</strong></td>
<td><strong>Fingerprint Scanning</strong> Software that allows a user to scan a fingerprint image and compare the digitized image data with fingerprints image data in a database.</td>
</tr>
<tr>
<td></td>
<td><strong>Iris Scanning</strong> Biometric identification by scanning the iris of the eye; Retinal recognition by means of scanning blood vessel patterns of the retina and the pattern of flecks on the iris.</td>
</tr>
<tr>
<td><strong>Pharmacy Automation</strong></td>
<td><strong>Automated Dispensing Machines</strong> A medication dispensing cabinet that automates the storing, dispensing and tracking of narcotics, floor stock and PRN medications in-patient care areas. Provides secure access to medications, while eliminating narcotic counts and keys. Interfaces with hospital ADT/billing systems to improve charge capture and materials management systems to track inventory.</td>
</tr>
<tr>
<td></td>
<td><strong>Carousels</strong> Physical devices that store day to day pharmaceutical supplies for manual or automatic picking of items for patient and nursing unit supply.</td>
</tr>
<tr>
<td></td>
<td><strong>Robotics</strong> Robotic technology used by pharmacies to conduct dispensing and cart fill functions and to deliver medications to medication cabinets for restocking.</td>
</tr>
</tbody>
</table>
Towards a Generalizable Time Expression Model for Temporal Reasoning in Clinical Notes

Sumithra Velupillai, PhD1,2, Danielle L. Mowery, PhD2, Samir Abdelrahman, PhD2,3, Lee Christensen, MS2, Wendy W Chapman, PhD2
1Department of Computer and Systems Sciences (DSV), Stockholm University, Stockholm, Sweden; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 3Computer Science Department, Faculty of Computers and Information, Cairo University, Egypt

Abstract

Accurate temporal identification and normalization is imperative for many biomedical and clinical tasks such as generating timelines and identifying phenotypes. A major natural language processing challenge is developing and evaluating a generalizable temporal modeling approach that performs well across corpora and institutions. Our long-term goal is to create such a model. We initiate our work on reaching this goal by focusing on temporal expression (TIMEX3) identification. We present a systematic approach to 1) generalize existing solutions for automated TIMEX3 span detection, and 2) assess similarities and differences by various instantiations of TIMEX3 models applied on separate clinical corpora. When evaluated on the 2012 i2b2 and the 2015 Clinical TempEval challenge corpora, our conclusion is that our approach is successful – we achieve competitive results for automated classification, and we identify similarities and differences in TIMEX3 modeling that will be informative in the development of a simplified, general temporal model.

Introduction

Accurate temporal identification and normalization is imperative for many biomedical and clinical tasks including generating timelines1, identifying phenotypes2, and creating problem lists3. Natural language processing (NLP) can help enrich event, phenotype, and problem detection by providing temporal context to event descriptions extracted from clinical texts. A major NLP challenge is developing and evaluating a generalizable temporal modeling approach that performs well across corpora and institutions. Few NLP system studies evaluate the generalizability of clinical temporal models and information extraction approaches that were developed for a particular use case or institution. To date, several NLP community challenges - 2015 Clinical TempEval4, based on the THYME corpus5, 2012 i2b26, and the 2014 ShARe/CLEF eHealth challenges7 - have attempted to encourage benchmarking and system evaluation for temporal extraction by applying adapted versions of the TimeML model8 to clinical texts of various report types from different institutions (Beth Israel Deaconess Medical Center, Partner’s Healthcare, and Mayo Clinic). The TimeML model enables automated temporal reasoning from text by defining a number of components, of which three components are central: events (e.g., diseases), time expressions (TIMEX3s, e.g., dates), and time relations (e.g. disease after date). Through this, it is possible to order events in time with respect to each other, and with respect to time expressions in the texts. Correctly identifying time expressions within a temporal reasoning system is necessary to capture specific mentions of time, and thus enable time anchoring and positioning of relevant events from narratives. These community challenges have established the state-of-the-art performance for NLP identification of these three core components in the adapted TimeML models.

For temporal modeling, overall performance has been published for each community challenge; however, there are few studies reporting system performance across different clinical corpora, and there has not been any in-depth analysis of performance on different temporal model subtypes. Our long-term goal is to create a generalizable temporal reasoning model that leverages and simplifies existing models. We initiate our work on reaching this goal by focusing only on TIMEX3 identification. In particular, we present a systematic approach to 1) generalize existing solutions for automated TIMEX3 span detection, and 2) assess similarities and differences in different instantiations of TIMEX3 models applied on different clinical corpora.

In the aforementioned challenges, successful approaches for TIMEX3 span detection, that is, correctly identifying and extracting specific mentions of time expressions in text such as “tomorrow”, “14 April 1998”, “for two weeks”, have either employed rule-based systems or machine learning approaches using lexical and syntactic features in combination with rule-based information. In the 2012 i2b2 challenge, the top performing approaches resulted in F1 scores between 0.8-0.91 for overlapping TIMEX3 spans9. The results in the 2014 ShARe/CLEF eHealth challenge
were lower: 0.37 overlap F1 \textsuperscript{1}. Our research lab, the Biomedical Language Understanding Laboratory (BluLab), recently participated in the 2015 Clinical TempEval challenge\textsuperscript{2}. Similar to other successful, high-performing NLP approaches, we developed a machine learning approach using local contextual features and information from a rule-based system to identify TIMEX3 spans. Our team ranked first on the TIMEX3 span detection task in this challenge, resulting in an overall exact F1 of 0.72. However, official scores in these challenges only report overall results for TIMEX3 detection. Moreover, to our knowledge, there are few studies reporting system performance across different clinical corpora, along with an in-depth analysis of similarities and differences in TIMEX3 annotations.

Methods
We performed our study on two corpora\textsuperscript{1}: the 2015 Clinical TempEval corpus, and the 2012 i2b2 challenge corpus. The Clinical TempEval corpus was created within the THYME project\textsuperscript{3} consisting of 440 clinical notes and pathology reports for colon cancer patients from the Mayo clinic (in total 293 documents for training, 147 for test). The i2b2 corpus consists of 310 discharge summaries (190 training, 120 test) from Partners Healthcare and the Beth Israel Deaconess Medical Center. We chose these two corpora for this study because they employ similar, but not identical, temporal models, and were specifically developed for temporal reasoning evaluation. Three main steps were employed: 1) TIMEX3 model mapping, with the aim of generating a common, comparable model, 2) state-of-the-art TIMEX3 classification assessment, with the aim of evaluating top performing TIMEX3 system approaches across corpora, and 3) qualitative error analysis, feature ablation evaluation and TIMEX3 type characterization, with the aim of completing an in-depth analysis.

Common TIMEX3 type model
In our attempt to characterize and analyze similarities and differences in separate instantiations of TIMEX3 models applied on different corpora, the first step was to create a common TIMEX3 type model. The two corpora employ slightly different TIMEX3 definitions and types. The Clinical TempEval corpus contains six TIMEX3 types: DATE, TIME, DURATION, PREPOSTEXP, QUANTIFIER and SET, in contrast to the i2b2 corpus which defines four types: DATE, TIME, DURATION and FREQUENCY. To enable comparison and analysis, we map the more fine-grained Clinical TempEval TIMEX3 types to the i2b2 types in the following way: PREPOSTEXP annotations are merged to DATE, and QUANTIFIER and SET annotations are merged to FREQUENCY. Examples are provided in Table 1. For our experiments, we maintained the training and test splits from each respective challenge corpus.

<table>
<thead>
<tr>
<th>TIMEX3 type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATE</td>
<td>05-04-1998, 29, February 2005, next week</td>
</tr>
<tr>
<td>PREPOSTEXP</td>
<td>Postoperative</td>
</tr>
<tr>
<td>DURATION</td>
<td>the first few days, the next 12 hours</td>
</tr>
<tr>
<td>FREQUENCY</td>
<td>several times a day, daily</td>
</tr>
<tr>
<td>QUANTIFIER</td>
<td>twice</td>
</tr>
<tr>
<td>SET</td>
<td>three times weekly</td>
</tr>
<tr>
<td>TIME</td>
<td>5.30 PM</td>
</tr>
</tbody>
</table>

Time expression extraction using ClearTK
To generalize existing solutions for automated TIMEX3 span detection, we applied the top-performing TIMEX3 span detection approach from the 2015 Clinical TempEval challenge\textsuperscript{9}. This is a UIMA-based machine learning solution using the ClearTK framework\textsuperscript{10}. A separate support vector machine (SVM: Liblinear) sequence label classifier (CleartkSequenceAnnotator) was created for each TIMEX3 type, where parameter setting (C-value) was determined by a grid search on a subset of the training data and set manually on the final model. The data was preprocessed with cTAKES v. 3.2 to extract lexical, morphological, and syntactic features\textsuperscript{11}. For the TIMEX3 types, gazetteer information specific to each type was also added as features, based on an adapted version of HeidelTime\textsuperscript{12}. A Begin-Inside-Outside chunking representation was used, with the following features: the current token, the current token with the two last characters stripped (\textit{decades} \rightarrow \textit{decad}), part-of-speech tag, numeric type (e.g. \textit{digit},

\textsuperscript{1} University of Utah Institutional Review Board approval was obtained for using these corpora for research
alphanumeric), capital type (e.g. all in upper case, mixed case), lower case, context tokens\(^2\), and whether or not the token is found in the TIMEX3 type-specific gazetteer. Compared to the successful approaches in the 2012 i2b2 challenge that used machine learning, this approach is similar, but not identical – for instance, Xu et al.\(^3\) employed the conditional random fields (CRF) algorithm and slightly different features. Roberts et al.\(^4\) used CRF for TIMEX3 boundary detection and SVM for type classification. Most other solutions were entirely rule-based\(^5\).

Three SVM models were trained using the datasets: 1) i2b2 training data, 2) Clinical TempEval training data, and 3), merged model with the two training data sets combined. System performance was evaluated on the i2b2 test data and the Clinical TempEval test data releases, using exact and overlapping precision, recall, and F1 as main outcome measures. We used the official evaluation script provided in the 2015 Clinical TempEval challenge\(^6\).

**Qualitative error analysis, feature ablation evaluation and TIMEX3 type characterization**

A qualitative error analysis to identify commonalities and differences was performed. We analyzed true positives on the results obtained by training a model on one corpus and evaluating on the other, to identify commonalities between the corpora and to characterize “core” TIMEX3 types. We also analyzed false positives and negatives, to identify and characterize differences between the annotations. Error analysis was performed on exact match only. Additionally, we performed a feature ablation study on the results obtained by this training/testing setup as well as on the results obtained when training a merged model with both corpora, evaluated on each test set. We built separate models with one feature removed each time, to identify which features seem to contribute the most for this task (i.e., if results are worse when removing one feature compared to using all features, this is an indication of an important feature).

**Results**

We present classification results on the i2b2 and Clinical TempEval test sets and results from the qualitative error analysis. First, we evaluate performance on within-corpus models. Second, we analyze performance on the two test sets on applied on the model trained on 1) i2b2 training data, 2) Clinical TempEval training data, and 3) merged training data. Finally, we present results from the qualitative error analysis, feature ablation evaluation and TIMEX3 type characterization, using the outputs from applying each test set on the model created from the other corpus’ training data. Overall corpus statistics are presented in Table 2.

**Table 2. TIMEX3 count (prevalence) statistics per type and corpus - Clinical TempEval and 2012 i2b2 datasets.**

<table>
<thead>
<tr>
<th>TIMEX3 Type</th>
<th>Clinical TempEval</th>
<th>i2b2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training data</td>
<td>Test data</td>
</tr>
<tr>
<td>DATE</td>
<td>2892 (75.7%)</td>
<td>1594 (76.7%)</td>
</tr>
<tr>
<td>DURATION</td>
<td>434 (11.4%)</td>
<td>200 (9.6%)</td>
</tr>
<tr>
<td>FREQUENCY</td>
<td>377 (9.8%)</td>
<td>225 (10.9%)</td>
</tr>
<tr>
<td>TIME</td>
<td>118 (3.1%)</td>
<td>59 (2.8%)</td>
</tr>
<tr>
<td>Total</td>
<td>3821 (100%)</td>
<td>2078 (100%)</td>
</tr>
</tbody>
</table>

**Temporal Information Extraction using ClearTK**

In Table 3, classification results on the two test datasets when applied on the model built from their corresponding training data are shown, strict and overlap. Results for the majority TIMEX3 type DATE are best overall. Results for DURATION and FREQUENCY are considerably higher when using overlapping evaluation criteria, in particular for DURATION, from 0.60 to 0.76 F1 for i2b2, and from 0.47 to 0.70 for Clinical TempEval. Precision is considerably higher for FREQUENCY when using overlapping evaluation, from 0.71 to 0.84 for i2b2 and 0.56 to 0.78 for Clinical TempEval. Results for the least frequent TIMEX3 type TIME differ between the corpora: for the

---

\(^2\) different context windows of surrounding tokens were employed for each TIMEX3 type as a result of experiments on the Clinical TempEval training data: 5 preceding, 3 following for DATE, 3/3 for TIME, 4/4 for DURATION and 5/3 for FREQUENCY

\(^3\) http://alt.qcri.org/semeval2015/task6/index.php?id=software
i2b2 corpus, precision is high when using overlap evaluation, but results are low overall for the Clinical TempEval corpus.

Table 3. Classification results for the mapped TIMEX3 types on the Clinical TempEval (CTE) and the i2b2 test datasets using the top-performing 2015 Clinical TempEval approach on the corresponding corpus training data. P=precision, R=recall.

<table>
<thead>
<tr>
<th>TIMEX3 type</th>
<th>strict - CTE/i2b2</th>
<th>overlap - CTE/i2b2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>span</td>
<td>0.80/0.79</td>
<td>0.66/0.70</td>
</tr>
<tr>
<td>class</td>
<td>0.78/0.77</td>
<td>0.65/0.69</td>
</tr>
<tr>
<td>DATE</td>
<td>0.82/0.80</td>
<td>0.76/0.78</td>
</tr>
<tr>
<td>DURATION</td>
<td>0.55/0.65</td>
<td>0.41/0.55</td>
</tr>
<tr>
<td>FREQUENCY</td>
<td>0.56/0.71</td>
<td>0.25/0.46</td>
</tr>
<tr>
<td>TIME</td>
<td>0.42/0.6</td>
<td>0.08/0.25</td>
</tr>
</tbody>
</table>

To further analyze differences and commonalities, the two test sets were also evaluated on all three created models. The least frequent TIMEX3 type (TIME) is not included since the number of training instances is insufficient. Not surprisingly, results are worst when applying each test set on the model created from the other corpus (e.g. training on Clinical TempEval and testing on i2b2), Figure 1. Specifically, results for DATE drop from 0.79 strict F1 (0.88 overlap F1) to 0.53 (0.61) on the i2b2 corpus, and from 0.79 (0.86) to 0.44 (0.57) on the Clinical TempEval corpus. For DURATION and FREQUENCY, the drop is from 0.60 (0.76) to 0.23 (0.44) and 0.56 (0.66) to 0.11 (0.18) on the i2b2 corpus. On the Clinical TempEval corpus, the drop is from 0.47 (0.70) to 0.31 (0.53) for DURATION and from 0.35 (0.49) to 0.06 (0.10) for FREQUENCY. Moreover, creating a merged model from the two training sets did not improve overall results for either test set (0.69 overall span strict F1 when evaluated on the i2b2 test set, and 0.69 when evaluated on the Clinical TempEval test set). However, results for DURATION on the Clinical TempEval test set are slightly improved when applied on the merged model (0.51 strict F1, 0.76 overlap).

Figure 1. Classification results per TIMEX3 type DATE, DURATION and FREQUENCY, strict and overlap. Three models: i2b2 model, Clinical TempEval (CTE) model, and merged (the two training sets combined) model, each evaluated on the i2b2 and CTE test sets.
**Qualitative error analysis, feature ablation evaluation and TIMEX3 type characterization**

We perform a qualitative error analysis on true positives, false negatives, and false positives (exact match) from the system output of each test set applied on the model created from the other corpus’ training data. This way, we are able to characterize corpus-specific annotations, and also commonalities between the two corpora.

**Clinical TempEval: 147 clinical and pathology notes, colon cancer (Mayo clinic)**

<table>
<thead>
<tr>
<th>TP</th>
<th>n=621</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATE: daily</td>
<td></td>
</tr>
<tr>
<td>DURATION: once</td>
<td></td>
</tr>
<tr>
<td>FREQUENCY: this time</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FN</th>
<th>n=651</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATE: today</td>
<td></td>
</tr>
<tr>
<td>DURATION: several years</td>
<td></td>
</tr>
<tr>
<td>FREQUENCY: twice a day</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FP</th>
<th>n=580</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATE: currently</td>
<td></td>
</tr>
<tr>
<td>DURATION: within the past x month</td>
<td></td>
</tr>
<tr>
<td>FREQUENCY: recent</td>
<td></td>
</tr>
</tbody>
</table>

| n=1167 |
| DATE: daily |
| DURATION: once |
| FREQUENCY: this time |

| n=1398 |
| DATE: recently |
| DURATION: overnight |
| FREQUENCY: per day |

| n=1109 |
| DATE: day of life x |
| DURATION: >48 hours |
| FREQUENCY: QID |

**i2b2: 120 discharge summaries (Partners Healthcare & Beth Israel Deaconess Medical Center)**

| n=621 |
| DATE: today |
| DURATION: several years |
| FREQUENCY: twice a day |

**Figure 2.** Venn diagram depicting error types and examples for each test set, based on system predictions for the i2b2 test set applied on the model built on the Clinical TempEval training data, and vice versa.

Most true positives on both test sets for DATE cover explicit dates such as “9 November, 2010”. These are also the most common DATE annotations in both corpora. However, DATE annotations such as “today”, “this time”, “many years ago”, “the same time” are also correctly classified in both test sets. For DURATION and FREQUENCY, common annotations include expressions like “several years” and “every x hours”. Examples are given in **Figure 2**.

However, the differences outweigh the commonalities between the corpora, as evidenced by the results. The Clinical TempEval DATE annotations contain a larger number of pre-/post-operative expressions that are not covered in the i2b2 corpus, as well as relative DATE instances such as “recent” and “now”. The i2b2 corpus, on the other hand, contains more DATE expressions typical for the document type (discharge summaries), such as “day of life x”, “hospital day x”, “day of delivery”, “day of transfer”, “the time of discharge”, that are not covered in the TempEval model. This is reflected also in the false positive results: when evaluating the i2b2 test data annotations on the Clinical TempEval model, expressions such as “currently”, “recent”, “intraoperatively”, “the future”, “the time”, “that time” are erroneously predicted as DATEs.

For DURATION, the i2b2 corpus contains expressions such as “>48 hours”, “the entire night”, “one months time”, “overnight”, “the next several hospital days” that are missed by the Clinical TempEval model, while the i2b2 model fails to represent Clinical TempEval DURATION annotations such as “quite a few years”, “a while”, “lifelong”, “overnight”.

1256
The i2b2 FREQUENCY annotations contain expressions not covered in the Clinical TempEval model, such as “q24”, “QID”, “qhs”, “bid”, “x 1”, while expressions such as “per day” and “xx/min” are common in the Clinical TempEval FREQUENCY annotations.

In addition to the lexical variations in the two corpora, the feature ablation analysis reveals that surrounding words are the most informative feature for all TIMEX3 types both when evaluating a trained model built on one corpus using the other test set, and when evaluating the merged model on each test set, Table 4. All other features have limited individual impact overall.

Table 4. Feature ablation study for error analysis. Each feature was removed when building the TIMEX3 model, keeping all remaining features. Results are presented for the TIMEX3 types DATE, DURATION and FREQUENCY, once when building a model on one corpus (Clinical TempEval (CTE) or i2b2) and evaluating on the other (cross-corpus), and once when building a model with both training sets (merged).

<table>
<thead>
<tr>
<th></th>
<th>DATE F1</th>
<th>DURATION F1</th>
<th>FREQUENCY F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cross-corpus CTE/i2b2</td>
<td>merged CTE/i2b2</td>
<td>cross-corpus CTE/i2b2</td>
</tr>
<tr>
<td>all features</td>
<td>0.53/0.44</td>
<td>0.77/0.76</td>
<td>0.23/0.31</td>
</tr>
<tr>
<td>-current token</td>
<td>0.48/0.48</td>
<td>0.73/0.69</td>
<td>0.18/0.23</td>
</tr>
<tr>
<td>-gazetteer</td>
<td>0.54/0.33</td>
<td>0.76/0.76</td>
<td>0.18/0.26</td>
</tr>
<tr>
<td>-pos</td>
<td>0.50/0.41</td>
<td>0.77/0.76</td>
<td>0.29/0.32</td>
</tr>
<tr>
<td>-context</td>
<td>0.38/0.16</td>
<td>0.58/0.46</td>
<td>0.00/0.04</td>
</tr>
<tr>
<td>-capitaltype</td>
<td>0.53/0.44</td>
<td>0.77/0.76</td>
<td>0.23/0.31</td>
</tr>
<tr>
<td>-numericitype</td>
<td>0.51/0.42</td>
<td>0.77/0.76</td>
<td>0.23/0.30</td>
</tr>
<tr>
<td>-lowercase</td>
<td>0.54/0.44</td>
<td>0.77/0.76</td>
<td>0.24/0.29</td>
</tr>
<tr>
<td>-ngram2</td>
<td>0.53/0.43</td>
<td>0.77/0.76</td>
<td>0.23/0.32</td>
</tr>
</tbody>
</table>

Discussion and Conclusion

In this study, we present a systematic, simplistic, and scalable approach to 1) generalize existing solutions for automated TIMEX3 span detection, and 2) assess similarities and differences in different instantiations of TIMEX3 models applied on different clinical corpora.

Retraining and applying the top-performing TIMEX3 span detection solution from the 2015 Clinical TempEval challenge on the i2b2 corpus produces competitive results: 0.74 strict F1 and 0.87 overlapping F1. The top ten performing systems in the 2012 i2b2 challenge resulted in strict F1 between 0.67-0.80 and overlapping F1 between 0.8-0.91. Combining the two training data sets to create a merged model slightly decreases overall results when evaluating on each test set, but results for DURATION on the Clinical TempEval test set are improved, indicating that the addition of new training data from a different corpus was informative.

Enabling characterization and analysis of commonalities and differences between different time expression instantiations by evaluating system output errors from a model trained on one corpus and evaluating on another is informative. In particular, this method reveals that some TIMEX3 expressions can be accurately covered across the two corpora, in particular for the majority type DATE. The main differences between the studied corpora can be categorized into two types: lexical (and document-type specific) variants (e.g. “qid”) and structural (document type). Further, the feature ablation study sheds light on the informativeness of individual features, where surrounding tokens provide context that is clearly crucial for improved performance.

Rule-based systems were successful in the TIMEX3 span extraction subtask of the 2012 i2b2 challenge. Such systems generally produce high recall, while machine learning-based systems generally produce high precision results. Given the new knowledge about lexical variants for different TIMEX3 types in different corpora, we also want to study how to incorporate this information and how to best make use of a combination approach for all elements needed in a temporal reasoning solution.

---

4 results for strict matching were recalculated using the official 2012 i2b2 evaluation script on system submission outputs provided by the challenge organizers
Contributions and limitations

The main contribution of our approach is that it is systematic and simple in acquiring and building a high-performing task model for TIMEX3 classification. Moreover, our results do not necessarily depend on the system approach for TIMEX3 detection - one could choose to apply any of the previous successful machine learning-based approaches for comparison or further development. Second, we achieve competitive results for automated classification on a new corpus. Third, we identify similarities and differences in TIMEX3 modeling which could be informative in the development of a simplified, general temporal model. Finally, our in-depth analysis identifies lexical variants of time expressions that could be useful for several clinical information extraction use-cases and feature engineering. However, we are planning to tackle the following potential future points:

Our proposed approach is evaluated only on two corpora annotated with similar temporal models. Other models have been proposed for the clinical narrative domain, such as the Temporal Constraint Structure\textsuperscript{15}. Applying our proposed approach on this model would probably require more mapping steps, as well as access to annotated documents for system development and assessment. An alternative approach to our mapping strategy could be to keep more granular types separate and build hierarchical structures (e.g. PREPOSTEXP is-a DATE) instead of mapping to broad TIMEX3 types – this way, one would get a richer representation while still keeping a generalizable overall model.

Document structure information, such as sections, is not taken into consideration when mapping annotations or for the feature engineering of the machine learning models, which is a central part of clinical narratives. We plan to incorporate this information in future extensions.

Our analysis of commonalities and differences is performed only on exact matched expressions: a richer overall picture would be obtained if also studying overlapping annotations. For successful unification of annotations, more training data, hierarchical ontologies, and annotation layers are needed, in particular for a more accurate and detailed analysis and model of less frequent types, such as FREQUENCY and TIME. The feature ablation study reveals that surrounding words are the most informative features. We intend to further study the impact of different context window sizes, and also whether the contextual information could be de-lexicalized in some way, to address generalizability across document types.

We will continue our work on developing a simplified, common temporal reasoning model for core elements needed in a full-scale temporal reasoning system. As a first step, we will employ our suggested approach on other available annotated corpora for further analysis and development. In particular, we plan to use the 2014 ShARE/CLEF eHealth corpus that contains disorder mention annotations along with semantic attributes, including time expression annotations. Since this corpus was created with a focus on disorder annotations, not overall temporal information, we expect that our proposed approach will have high coverage on extracting relevant time expressions, but that tailoring will be necessary for disorder context. We will also extend our work to include temporal expression normalization, as well as event span and attribute classification, and temporal relations.

Acknowledgements

We thank the Mayo clinic and the challenge organizers for providing access to the 2015 Clinical TempEval corpus, and i2b2 and the 2012 challenge organizers for providing access to their corpus. This work was partially funded by Swedish Research Council (350-2012-6658), NLM R01 LM010964, and United States Department of Veteran Affairs CREATE Information Extraction and Visualization Toolkit (IE-Viz) CRE 12-312.

References


Organizational Uses of Health Information Exchange to Change Cost and Utilization Outcomes: A Typology from a Multi-Site Qualitative Analysis

Joshua R Vest, PhD, MPH1 and Erika Abramson, MD, MS1,2

1Division of Health Policy and Economics, Department of Healthcare Policy and Research, Weill Cornell Medical College, New York NY; 2Department of Pediatrics, Weill Cornell Medical College, New York, NY

ABSTRACT

Health information exchange (HIE) systems facilitate access to patient information for a variety of health care organizations, end users, and clinical and organizational goals. While a complex intervention, organizations’ usage of HIE is often conceptualized and measured narrowly. We sought to provide greater specificity to the concept of HIE as an intervention by formulating a typology of organizational HIE usage. We interviewed representatives of a regional health information organization and health care organizations actively using HIE information to change patient utilization and costs. The resultant typology includes three dimensions: user role, usage initiation, and patient set. This approach to categorizing how health care organizations are actually applying HIE information to clinical and business tasks provides greater clarity about HIE as an intervention and helps elucidate the conceptual linkage between HIE and organizational and patient outcomes.

INTRODUCTION

Health information exchange (HIE) is the electronic sharing of patient information between different health care organizations and providers.[1] Through HIE, physicians, clinical staff, and administrative users have access to patients’ tests, reports, encounters, and summary documents from other providers. Improved access to more comprehensive information may support decision-making, inform providers of additional medications or allergies, and help avoid repeated or duplicate testing.[2] The literature on the effectiveness of HIE is mixed,[3,4] but several evaluations have suggested HIE usage can help avoid unnecessary utilization[5,6] and reduce health care utilization and costs.[7-9] These evaluations tend to either view HIE as a structural capability of the health care organization,[e.g. 10,11,12] or examine the effects of accessing HIE systems in a specific clinical context.[e.g. 5,7,9] The use case underlying many of these evaluations is that HIE fills in the gaps in patient information for clinicians.[2,13]

However, HIE is a much more complex intervention in the daily operation and practices of health care organizations. For example, the use cases for HIE extend to non-clinical staff in health care organizations, such as case managers or social workers, and HIE can support non-clinical functions.[14] Also, conceptualizing HIE activity as a uniform structural resource for an organization obscures much of the variation in information access and usage. Users of HIE systems retrieve and receive different types of information[15] and in different formats (e.g. reports, push messages, or online queries).[16] Furthermore, integrating information from HIE into clinical care or business processes requires more steps than is reflected by basic measures of system access.[17] With all these known variations, how organizations actually put HIE delivered information to use to improve patient outcomes and reduce costs is not well defined.

We sought to provide greater specificity to the concept of HIE as an intervention to change cost and utilization outcomes. While evaluations according to the availability or access of HIE systems have been insightful, they do not capture how individuals and health organizations actually act upon HIE information or integrate HIE into existing workflows and clinical practices. Using qualitative methods, we examined the basic question, how are organizations using the community-wide patient information made available through HIE? To answer this question, we developed a typology of the organizational applications of HIE. This typology provides greater clarity about HIE as an intervention and facilitates the conceptual linkage of this technological intervention to patient outcomes and organizational changes.
METHODS

We selected a qualitative approach to capture the variety of potential patient populations, clinical care settings, work processes, and applications of HIE.

Study site
All data were collected from health care organizations working with the Bronx Regional Health Information Organization (RHIO) in New York. The RHIO is a non-profit organization that facilitates HIE for more than 60 inpatient and ambulatory care organizations and includes information on more than 550,000 patients. The Bronx RHIO was established in 2005 and offers notification alerts of patient utilization, DIRECT messaging, a portal to query patient information, a centralized data repository for analytics, electronic referral management, and analytic services.[18] In 2012, the Bronx RHIO was awarded a CMS Health Care Innovation Award to develop data registries and expand analytic capabilities in order to reduce the cost of care for Medicaid and Medicare patients in the borough. This expansion of services makes the Bronx RHIO an ideal setting for a typology development as it resulted in a large and diverse number of activities supported by HIE. As part of the grant, the Bronx RHIO customizes services for each participant through collaborative consolation, so that HIE information supports each organization’s specific portfolio of services or interventions. This analysis was conducted as part of broader evaluation that includes a quantitative evaluation of impact and costs.

Selection of participants
We conducted interviews with 25 health care organization leaders, clinicians and employees, as well as representatives from the Bronx RHIO from June 2014 to March 2015. Interviewees had titles such as: CMIO, Section Chief, Director, Nurse, Case Manager, Physician, Statistician, and Patient Navigator. A total of 5 organizations were included. We purposefully sought interviewees who were knowledgeable about the organization’s work with the Bronx RHIO, users of information obtained by HIE, and who were participating in the Bronx RHIO’s current CMS supported HIE activities. Bronx RHIO staff identified the key contacts and the key contacts helped to identify the relevant individuals within the organization. Consent was obtained from each individual interviewee. Interviews also occurred opportunistically during the data collection process. Recruitment stopped once all community partners using HIE to support their activities were represented and thematic saturation had been reached. All interviews were audio recorded and transcribed.

Interviews
Interviews followed a semi-structured format with open-ended questions. The interview guide asked about the characteristics of the organization and specifics about the activities (including day to day activities, interventions or projects) supported by HIE. Questions about the role of HIE in these activities were guided by the Triangle Model for evaluating health information technology interventions, which considers technology usage as the product of patient, provider, technology and organization interaction.[19] Data collection occurred through a combination of on-site (72%) and telephone interviews. Interviews lasted an average of 33 minutes.

Interviews occurred in three distinct steps. First, we conducted a focus group with representatives of the Bronx RHIO to obtain an overview of all partner activities. The focus group included a group card-sorting activity, where the respondents were asked to group the partners’ activities according to: 1) project goals; 2) how information is accessed; 3) who accesses the information; and 4) no set criteria (free choice). Next, we interviewed the leadership of the participating health care organizations. Interviews occurred only after the organization was actively applying HIE to an activity (i.e. the site had “gone live”). Organizations were active once they: 1) established procedures and policies for Bronx RHIO information usage; 2) HIE information was being accessed as part of the activity; and 3) HIE information had been applied in the delivery of health care services to patients. Finally, we interviewed those individuals using HIE information or involved in the activities supported by HIE. This approach limited our sample to active users and excluded any organization planning or simply “piloting” HIE usage.

Analysis
The first objective with our qualitative analysis was to develop a typology of HIE usage. We did not select any particular theoretical lens for analyzing the data, but instead followed an iterative, general inductive and comparative approach.[20] First, through a joint reading of three transcripts and review of the card sorting exercise, we identified preliminary themes relevant to defining our typology using open coding. Subsequently, we coded all transcripts together during joint reading sessions, continually checking, revisiting, and adjusting our typology in light of
additional data. As a member check, we reviewed our typology with Bronx RHIO representatives and other HIE evaluators. The second objective was to identify any factors that acted as barriers or enablers of successful usage of HIE. During interviews, we did not formally define success, but for our analyses we defined success solely as a reported application of HIE to meet a goal or job requirement. Therefore, success was both self-reported and could vary between job types and organizations. Determinants of success were identified concurrently during our analysis using an open coding approach. Through discussion and regular meeting sessions, synonymous codes were merged and we grouped each identified theme into the broader categories of determinants suggested by the Triangle Model (i.e. axial coding). Data analysis used NVivo10. The study was approved by the Weill Cornell Medical College Institutional Review Board.

RESULTS

Interviewees represented specialty care centers at an integrated hospital system, a hospital emergency department, a federally qualified health center (FQHC), and a case management organization. We identified 12 examples of different applications of HIE during interviews (Table 1). Some instances of usage were repeated across the organizations (e.g. most had staff authorized to access the Bronx RHIO’s standalone web-based portal). However, other organizational activities were supported by reports consisting of line lists of patients or by automatic notifications of patient emergency department utilization and hospital admissions.

Table 1. Organizational applications of health information exchange – active users of the Bronx Regional Health Information Organization during 2014-2015

<table>
<thead>
<tr>
<th>Organization</th>
<th>Description of HIE-supported project or activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital</td>
<td>Listing of patient eligible for care management due to encounters at other facilities</td>
</tr>
<tr>
<td>Health system</td>
<td>Query-based exchange to support geriatric care in the emergency department</td>
</tr>
<tr>
<td>Health system</td>
<td>Listing of recent emergency encounters and admissions for scheduled patients</td>
</tr>
<tr>
<td>Health system</td>
<td>Listing of recently discharged patients to enroll in specialty care centers</td>
</tr>
<tr>
<td>Health system</td>
<td>Listing of health home patients hospitalized for patient navigators to contact</td>
</tr>
<tr>
<td>FQHC</td>
<td>Listing of patients in need of asthma follow-up after emergency visit</td>
</tr>
<tr>
<td>FQHC</td>
<td>Listing of demographic and contact information for health home assigned patients</td>
</tr>
<tr>
<td>FQHC</td>
<td>Query-based exchange during patient visit</td>
</tr>
<tr>
<td>FQHC</td>
<td>Automated notifications of patients’ emergency encounters and admissions</td>
</tr>
<tr>
<td>Case management</td>
<td>Automated notifications of patients’ emergency encounters and admissions</td>
</tr>
<tr>
<td>Case management</td>
<td>Listing of demographic and contact information for care management assigned patients</td>
</tr>
<tr>
<td>Case management</td>
<td>Listing of patients’ recent emergency visits and discharges</td>
</tr>
</tbody>
</table>

Typology

Our typology includes three independent dimensions: USER ROLE, USAGE INITIATION, and PATIENT SET (see Figure 1 on the next page). The typology is applied at the level of the application of HIE in the context of a specific program or activity.

The primary, and most distinguishing, characteristic of each of the organizational applications of HIE was the USER ROLE. Formally, we defined the dimension as the “ultimate consumer of the information.” The USER ROLE dimension does not necessarily measure the individual who interacted with the HIE systems, but instead the focus is on the individual within the organization who was expected to act upon the information or for whom the information was supposed to support their work. This distinction reflects the features we observed during interviews. First, information obtained via HIE could have been delivered to an intended user in any fashion (e.g. a portal, notification, or report). Second, the HIE information may effectively pass through many different hands on the way to the “ultimate consumer” (e.g. proxy users, or analysts who produce reports to give to other staff). We subdivided USER ROLE into two categories. Clinical users included all physicians, nurses, and case managers applying HIE information to direct patient care. Administrative users were managers, analysts, and staff not engaged in clinical care.
The second dimension of the typology describes how information was obtained from HIE. Instead of focusing a specific type of technology used (e.g. DIRECT compared to a portal), we focused on a larger concept: who was responsible for obtaining the information for the end user or who initiates usage (i.e. USAGE INITIATION)? In our sample, the organizations either relied on the Bronx RHIO staff to provide the information in a form for use by end users (i.e. externally generated) or the responsibility for ensuring end users had access to information was managed internally. For example, the Bronx RHIO staff would electronically send a detailed report of their clients’ recent emergency encounters and utilization history. To the organization, the information was created externally. In contrast, some organizations utilized staff and resources to generate similar reports or to query the centralized data repository for patient information. As a case in point, one nurse case manager reported she relied on the Hospital’s statistician “to generate that list of potential [clients]” for her. These were instances of internally initiated access to HIE information.

The final dimension, PATIENT SET, refers to the type of patient to whom the information is being applied. Organizations were applying HIE information to two broad types of patients. First, they were using HIE to find information on a pre-defined set of patients, in other words on patients that were already seeking ongoing care from the organization. Generally, these patients were selected subsets of the entire patient populations defined by a specific clinical condition (e.g. diabetes, asthma, etc.) or by utilization levels (e.g. frequent emergency department users). For example, the FQHC was using HIE information to identify which of their existing diabetes patients were out of date on laboratory testing. The Health System applied HIE in a similar fashion by generating line listings of their geriatric and asthma clinic patients’ recent emergency department visits and admissions. Alternative to using HIE to find information about existing patients was applying a set of criteria in order to find patients for inclusion in program or intervention. The criteria for finding patients was often a specific condition or response to a type of event. As an example of this type of patient finding, one organization used HIE to identify patients to include in their medical home and another used HIE to supplement missing demographic data in order to contact individuals recommended by the state Medicaid program for care management. As a third example, the Hospital and Health System used HIE to identify patients that had been recently discharged without usual sources of care to recruit for their care management or outpatient programs.

Determinants of successful application of HIE
As noted, all organizations were actively using HIE information within the organization. The degree of success was both varied and dependent upon multiple factors. ORGANIZATIONAL FACTORS, such as available resources, established procedures, and dedicated staff, were supportive for both administrative and clinical users. For example, the Health System queried the central data repository to generate a “lost to follow up report” to identify patients recently discharged who had not been seen at their specialty clinics. To act on this information, the Health System had dedicated staff and a “call center does exist, and they can and do respond to orders to make those calls.”

Figure 1. Multi-dimensional typology of organizational applications of health information exchange (HIE)
Likewise, for clinical users, the FQHC structured work processes so that information retrieval from HIE was the responsibility of medical assistants and not other clinicians. As another example, the Case Management organization has staff and procedure in place to respond to any automated notifications that are received after normal business hours.

Generally, the application of HIE fit better into administrative workflows than for PROVIDERS. Physicians and nurse practitioners were not often active users of the query-based HIE, and HIE information was more likely to get in front of clinical users if work practices made sure the information was delivered to them or if another user queried databases on their behalf. For physicians, there was variable utility in knowing about recent discharges or emergency department visits. However, administrative users could use even a small bit of information (such as a telephone number), and even outside the context of a patient visit, to perform outreach or other activities. There was generally more certainty that the administrators would act on the information than clinicians. For example, as a manager noted, “I can control the [administrative staff]. They work for me. The doctors don’t…”

PATIENT FACTORS could inhibit the effectiveness of the organization’s application of HIE. As an example, each morning clinicians in specialty clinics serving diabetic and asthmatic patients were provided lists of the recent emergency encounters and admissions for that day’s scheduled patients. Clinicians found that information useful, but the utility of the information was limited because “my no show rate is so high that many of those patients on that report probably don't show up for their appointments.” In particular, asthma patients were not inclined to follow up with a visit after an emergency department encounter because their symptoms were controlled or their acute needs had already been addressed. In contrast, the show rate for geriatric patients was much higher and as a result the information from HIE was “a great tool.” This may be due to the fact that geriatric patients tended to be quite loyal to their primary care providers who were coordinating the entirety of their care.

Unsurprisingly, lack of integration between HIE TECHNOLOGY and other systems was often viewed as a barrier to usage. This was most noted for query-based HIE, because none of the organizations had a single sign on option within their electronic health record (EHR). As a result, usage of query-based HIE suffered. As one interviewee noted: “I don't actually go to the [portal] because it's very cumbersome….I mean, so how does it fit into my day? It doesn't fit into my day. …” In contrast, internally initiated reports from the centralized analytics database by report writers or data analysts fit much better into workflows. Likewise, interviewees reported alert notifications and aggregated reports created by the Bronx RHIO fit easily into workflows and were much easier approaches to use HIE information.

Finally, patient CONSENT was a recurring theme in our interviews and was consistently a barrier. New York State policy requires active patient consent and that consent be obtained at each site. Clinical and administrative users were unable to access information they believed they needed for unconsented patients. Organizations varied in their approaches to dealing with the challenge of consent, but most tried to incorporate the consent paperwork as part of the registration process. The FQHC had a community health worker on site (funded by the Bronx RHIO) solely to consent and educate patients. This approach was particularly successful, because it coupled a culturally competent representative with technology and a standardized work process. If the community health worker could not obtain patient consent, the clinician was notified within the EHR to also try to obtain consent.

LIMITATIONS

This typology and the factors identified with HIE usage are limited by several factors. First, our findings are based on a single RHIO’s line of services. However, the number of services offered by the RHIO is broad and to a diverse set of health care organizations. Furthermore, these services are consistent with our broader experiences with other HIE-facilitating organizations. Second, it is possible to conceptualize organization’s application of HIE on additional dimensions not represented in our typology. For example, our typology does not include the information content, which has been applied in prior HIE studies of usage. We choose to keep the number of dimensions limited to make it more manageable and easier to apply generally. Also, the relevance of clinical and demographic information as a distinguishing feature of usage is not always clear. Both are often needed and both are often present from any type of usage of HIE (i.e. push and pull systems both provide access to clinical and demographic information). Likewise, the Triangle Model proved to be a useful approach for categorizing themes identified from different care settings, but it is possible that basing our interview guide on an alternative framework would have resulted in additional determinants. Additionally, our typology reflects organizational activity relatively early in the
course of their HIE usage, as such we did not observe all possible combinations of our USER ROLE – INITIATES
USAGE - PATIENT SET typology (as some may not occur frequently). We plan on conducting a second wave of
interviews to better understand how HIE activities evolve over time. Lastly, we did not independently assess or
define successful use of HIE, but let the interviewees define what they viewed as a “successful application of HIE.”

DISCUSSION

Based on interviews with those facilitating HIE activities and health care professionals using HIE services, we
developed a three dimensional typology of organizational use of HIE. Each instance of an organization’s application
of HIE can be categorized according to the type of USER ROLE (clinical or administrative), the USAGE INITIATION
(internally or external), and the PATIENT SET (pre-defined patients or patients to be found). This approach to
categorizing how health care organizations are actually applying HIE information to clinical and business tasks
provides greater clarity about HIE as an intervention and helps elucidate the conceptual linkage between HIE an
organizational and patient outcomes.

Importantly, this study and the resulting typology, illustrates the broad set of possible applications of HIE to health
care organizations. Generally, HIE has been conceptualized and evaluated as a clinical intervention. As
reported here, HIE supported direct clinical care, but also case management activities and patient enrollment into
specific programs. These additional use cases are of potential value to health care organizations and also illustrate
the diverse types of users associated with HIE. Other evaluations have noted that HIE is used by more than just
physicians, but this typology helps explain the broader purposes and activities of these non-clinical users. In
similar fashion, the typology highlights the unique role HIE can fulfill for health care organizations: as a source of
information on the unfamiliar, hard to reach, and sought out type of patient. Each interviewee had multiple
information systems supplied by their organization or even partner entities (like the state Medicaid agency). Once
those sources had been exhausted, interviewees were turning to HIE to fulfill their information needs.

Considering how organizations actually apply HIE to the care of their patients may help explain many of the current
research findings and can provide guidance to future studies. For example, several evaluations have viewed HIE
from the adoption / implementation perspective by measuring HIE simply as present or absent. This is a perfectly
valid approach and very much in line with the idea of information systems as structural component of the
organization, but findings from the literature using this measurement approach are mixed. Our
typology can help move measurement a step closer to actual processes. Instead of measuring HIE as a structural
feature only, our typology characterizes and clarifies the underlying processes organizations use to apply the
information obtained via HIE. This potential tighter coupling could help better inform the types of outcomes we
should expect from HIE interventions and also suggests considering potential organizational and operational
outcomes in addition to the utilization outcomes which are the focus of most of the existing research. For example,
organizations obtaining regular HIE reports containing information on new patients to enroll in a program clearly
fits expectations of efficiency and productivity gains. While these are important gains, it would be hard pressed to
expect quick changes to patient care and quality from such a usage strategy. Alternatively, changes to care
utilization and cost reduction might be more logical for clinical users with on-demand access to information working
to intervene on a set of pre-defined high-risk patients. Both use cases have a value and utility for a high performing,
cost conscious organization. However, these potential impact of administrative and clinical HIE use cases most
likely will not be reflected in the same dependent variables.

For health care managers, the immediate impact of adoption HIE might be most quickly realized in administrative,
patient finding projects. For one, the risk of non-acceptance of technology tends to run higher among clinicians and
physicians in particular. Second, implementing reports or queries into administrative workflow might be easier.
Third, this type of usage does not require the level of resources or staffing necessary to, for example, mount an
effective case management program to respond to notification alerts. Importantly, as the instances of usage
described in this study illustrate, administrative usage requires minimal information like addresses or telephone
numbers to be effective in reaching previously lost patients. If the patients were truly lost to follow up or impossible
to reach with existing information, any additional patient finding using HIE information is a gain that would not be
realized otherwise. Although not a particularly exciting use case, the 30 day readmission reduction program and
population health approaches like Accountable Care Organizations have heightened the importance of this basic
activity. Administrative use cases may not necessarily be what health information technology advocates and health
reforms hope for from HIE, but it is a start. It then becomes incumbent on the organization and the HIE facilitating
organization to make sure that HIE does not get regulated to only as a back-office application, but that the clinical use cases are adopted as well.

Finally, this study reiterates that health care organizations and the usage of health information technology are still subject to the influences of the external environment. In this study, patient consent was a barrier to effective usage. These policies governing the rules and requirements for obtaining patient consent are outside of health care organizations’ or the RHIO’s direct control. In New York State, consent policies are set statewide and developed through a collaborative process facilitated by the New York eHealth Collaborative and includes the State Department of Health, provider representatives and the public.[24,25] Organizations facing externally defined constraints have basically three broad options. The first is to try to address the challenge internally by shifting work process to meet the externally defined expectations. The Bronx RHIO’s funding of onsite staff and organizations building in the consent process into workflows were effectively this approach. The goal of both of these approaches was only to mitigate the perceived challenges dictated by the environment, e.g. New York’s active consent policy places responsibility for consent on the RHIOs, but the actual operationalization of the consent process occurs within health organizations. A second option is to attempt to change the external environment. Health care organizations in states with opt-in policies face workflow challenges and have to expend resources on a process that a) is the benefit of the patient and b) is a burden that their competitors and colleagues in opt-out states do not have to absorb. Health care organizations in states with opt-in policies that are serious about HIE a valuable part of their technology strategy should engage health policy makers to facilitate access to necessary information in a way that still meets patients expectations for privacy. Health policy makers should be willing to work with health care organizations on this issue, because the risk is if they don’t health care organizations will adopt the third potential response to external pressure: HIE just won’t be used.

CONCLUSION

Our study creates a novel typology for characterizing how health care organizations are actually applying HIE information to clinical and administrative tasks. Our typology consists of three distinct dimensions: USER ROLE, USAGE INITIATION, and PATIENT SET. Use of this typology can help support future evaluations and implementations by focusing on the distinct role of the information obtained from HIE plays in activities desired to improve organizational and patient outcomes.

ACKNOWLEDGEMENTS

We thank the interviewees and their respective organizations for their time and cooperation. Also, we thank Drs. Rainu Kaushal, Lisa Kern, and Zachary Grinspan for their insights and assistance. This publication was made possible by Grant Number 1C1CMS331065 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services.

The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services or any of its agencies.

REFERENCES


A Graph Based Methodology for Temporal Signature Identification from EHR

Fei Wang¹, Ph.D., Chuanren Liu², Yajuan Wang³, Ph.D., Jianying Hu³, Ph.D., Guoqiang Yu⁴, Ph.D.
¹Department of Computer Science and Engineering, University of Connecticut
²Decision Sciences & MIS Department, Drexel University
³Healthcare Analytics Research Group. IBM T. J. Watson Research Center
⁴Department of Electrical and Computer Engineering, Virginia Tech
Email: fei.wang@uconn.edu, liuchuanren@gmail.com, {wangya, jyhu}@us.ibm.com, yug@vt.edu

Abstract

Data driven technology is believed to be a promising technique for transforming the current status of healthcare. Electronic Health Records (EHR) is one of the main carriers for conducting the data driven healthcare research, where the goal is to derive insights from healthcare data and utilize such insights to improve the quality of care delivery. Due to the progression nature of human disease, one important aspect for analyzing healthcare data is temporality, which suggests the temporal relationships among different healthcare events and how their values evolve over time. Sequential pattern mining is a popular tool to extract time-invariant patterns from discrete sequences and has been applied in analyzing EHR before. However, due to the complexity of EHR, those approaches usually suffers from the pattern explosion problem, which means that a huge number of patterns will be detected with improper setting of the support threshold. To address this challenge, in this paper, we develop a novel representation, namely the temporal graph, for event sequences like EHR, wherein the nodes are medical events and the edges indicate the temporal relationships among those events in patient EHRs. Based on the temporal graph representation, we further develop an approach for temporal signature identification to identify the most significant and interpretable graph bases as temporal signatures, and the expressing coefficients can be treated as the embeddings of the patients in such temporal signature space. Our temporal signature identification framework is also flexible to incorporate semi-supervised/supervised information. We validate our framework on two real-world tasks. One is predicting the onset risk of heart failure. The other is predicting the risk of heart failure related hospitalization for patients with COPD pre-condition. Our results show that the prediction performance in both tasks can be improved by the proposed approaches.

1 Introduction

Patient Electronic Health Records (EHRs) [6] is one of the major carriers for conducting data driven healthcare research. There are various challenges if we work directly with EHRs, such as sparsity, noisiness, heterogeneity, bias, etc [5]. One important aspect for mining EHR is how to explore the temporal relationships among different medical events within patient EHRs. Many approaches have been proposed for temporal mining of EHRs. For example, Lasko et al. [11] proposed a deep learning method for obtaining patterns from lab value signals. Kale et al. [7] applied deep learning to discover the physiomes from the physiological streams obtained in Pediatric Intensive Care Unit (PICU). Zhou et al. [24] proposed an optimization based technology for discovering the feature groups within which the raw medical features have similar evolving patterns. For all these works, the authors either identified as some evolving pattern on the values of a specific medical feature (e.g., lab test or physiological stream), or a group of medical features (e.g., diagnosis, medication or both). Another important type of feature is the temporal pattern across different medical features, which is referred to as temporal signatures in this paper. The existing approaches on temporal signature identification are mostly based on sequential pattern mining [2, 18] or temporal abstraction [17, 19, 20]. One major challenge for these methodologies working on EHRs is pattern explosion, which is the phenomenon that too many patterns are identified from the patient EMR corpus with improper support threshold. One could try to solve this problem by increasing the support threshold value used by sequential pattern mining but the mined signatures are then typically trivial. Therefore there is an urgent need on an effective way to identify a reasonable number of clinical meaningful signatures.

One cause of pattern explosion for traditional approaches is the sequence based representation: lots of the time the
finest time granularity of patient EHR is day, within each day many events can happen. This generates a huge number
of sequential patterns when the support threshold is relatively lower [2, 18]. In this paper, we propose a novel graph
based representation for patient EHRs. We represent the EHRs of every patient as a graph. The nodes in the graph are
the medical events (i.e., diagnosis, medications, lab tests, etc.). The edges indicate the temporal relationships among
the events in the EHRs of the corresponding patient. Every edge points from an event to another event that took place
later in time. A weight will also be associated with each edge, which encodes the average duration between the two
events in EHRs. A bases learning framework is then developed to identify the bases that can be used to compose all
those temporal graphs. We present several concrete instantiations of such framework and validate its effectiveness on
a real-world EHR data warehouse both quantitatively and qualitatively.

It is worthwhile to highlight the following aspects of the proposed graph based framework:

- Comparing to the traditional sequence based representation, temporal graph is more compact, which makes the
downstream analytics procedure more efficient.
- With graph based representation, the detected signatures are subgraphs instead of subsequences. Each subgraph
is composed of a set of subsequences. This effectively alleviates the pattern explosion problem while at the
same time retains the interpretability of the mined signatures.
- We provide concrete instantiation examples of the proposed framework in a completely unsupervised scenario,
as well as in scenarios where we incorporate expertise knowledge as regularizers.
- The framework is validated on a real-world EHR data warehouse with two specific clinical scenarios. One
is early detection of Congestive Heart Failure (CHF). The other is hospital readmission prediction of CHF
patients with Chronic Obstructive Pulmonary Disease (COPD) pre-conditions. We showed that the prediction
performance can be improved with the mined temporal signatures.

2 Related Work

This section reviews the existing work that are closely related to the research proposed in this paper. One is electronic
phenotyping approaches. The other is temporal knowledge extraction.

2.1 Electronic Phenotyping

Electronic phenotyping refers to the process of identifying phenotypes from patient EHRs, which is the procedure of
extracting clinically relevant features. There are quite a few existing electronic phenotyping works. For example, Ho
et al. [3, 4] formulates the patient EHRs as tensors, wherein every mode represents a specific type of medical event.
The entries in the tensor encode the interaction of those features (e.g., the frequency of a medication and a primary
diagnosis). Then they proposed a tensor factorization based approach for identification of the phenotypes. Zhou et al.
[24] formulates EHRs as temporal matrices with medical events as one dimension and time as the other dimension.
They propose an optimization based technology for discovering the phenotypes within which the raw medical features
have similar time-evolving patterns. Lasko et al. [11] proposed a deep learning method for obtaining phenotypes
from continuous lab value signals, where they first adopted Gaussian process regression to impute the missing lab test
values. Kale et al. [7] applied deep learning to discover the physiomes from the physiological streams obtained in
Pediatric Intensive Care Unit (PICU). For all these works, they either define a phenotype as some evolving pattern on
the values of a specific medical feature (e.g., lab test or physiological stream), or a group of medical features (e.g.,
diagnosis, medication or both). They did not consider the temporal relationships across different medical events, which
could be crucial as they suggest important information on the impending disease conditions.

2.2 Temporal Knowledge Representation

Knowledge representation from temporal data is a hot research topic in both data analytics and biomedical infor-
matics. For continuous time data, one popular approach is to transform them into discrete symbolic representations
(string, nominal, categorical, and item sets). Popular approaches include Piecewise Linear Approximation (PLA) [10], Adaptive Piecewise Constant Approximation (APCA) [8], Symbolic Aggregate approXimation (SAX) [13], Piecewise Aggregate Approximation (PAA) [9], etc. One can refer to [13] for a survey on these approaches.

For discrete time data, Mörchen et al. [15, 16] proposed the Time Series Knowledge Representation (TSKR) as a pattern language (grammar) for temporal knowledge discovery from multivariate time series and symbolic interval data, where the temporal knowledge representation is in the form of symbolic languages and grammars that have been formulated as a means to perform intelligent reasoning and inference from time-dependent event sequences. More recently, Wang et al. [22] proposed a convolutional framework to extract temporal signatures in discrete time data using the Temporal Event Matrix Representation (TEMR), which is shown to have wide applicability to a variety of data and application domains that involve large-scale longitudinal data.

The temporal graph we propose in this paper provides an alternative way to represent the temporal knowledge contained in discrete time data. The temporal graphs capture temporal structures hidden in the sequences in a more compact way, where the nodes in the graph are events appeared in the EHR and the directed edges encode the temporal relationships between pairwise events. In this representation, the events missing in patient EHRs will not appear in the graph, and the repeated pairwise events with the same ordering will only appear once in the graph. With this representation, the temporal graph is robust and resistant to sparse, noisy, and irregular observations. Moreover, this representation is very intuitive to interpret the temporal relationships among different medical events in patient EHRs. Another advantage is that with graph based representation, the detected phenotypes (or patterns) will also be in the form of graphs, which can be viewed as a nature aggregation of sequential patterns. In this way, we can effectively alleviate the pattern explosion problem.

3 Methodology

In this section we will introduce the details of our temporal graph based framework for phenotype identification from patient EHRs. First we present the basic definition of temporal graph and how it is constructed.

3.1 Temporal graph construction

Suppose we have a set of event sequences $S = \{s_n : n = 1, \cdots, N\}$ where $N$ is the number of sequences. Each event sequence is denoted by $s_n = ((x_{nl}, t_{nl}) : l = 1, \cdots, L_n)$ where $L_n$ is the length of $s_n$. In other words, we observe event $x_{nl}$ at time $t_{nl}$ in the sequence $s_n$. We order the events $x_{nl} \in \{1, \cdots, M\}$ so that $t_{np} \leq t_{nq}$, for all $p < q$. Figure 1 illustrates an example of such graph.

With the observed event sequences, we construct the following temporal graph for each sequence $s_n$:

**Definition 1 (Temporal graph)** The temporal graph $G^n$ of sequence $s_n$ is a directed and weighted graph with our event set as its node set $\{1, \cdots, M\}$, where the weight of the edge from node $i$ to node $j$ is defined as

$$W^n_{ij} = \frac{1}{L_n} \sum_{1 \leq p \leq q \leq L_n} \left[ x_{np} = i \land x_{nq} = j \right] \kappa \left( t_{nq} - t_{np} \right),$$

where $\kappa(\cdot)$ is a non-increasing function.
\[ \kappa(\cdot) \text{ is a non-increasing function defined as follows, thus the closer event } i \text{ and event } j \text{ appear in } s_n, \text{ the higher } W_{ij}^n \text{ will be.} \]

\[
\kappa(\delta) = \begin{cases} 
\exp(-\delta/r) & \delta \leq \Delta \\
0 & \delta > \Delta 
\end{cases}
\] (2)

Here \( \Delta \) is a time horizon so that we only create an edge between a pair of events if the duration between them is smaller than or equal to \( \Delta \). \( r \) controls the locality of the edge computation in the temporal graph. Namely, a larger \( r \) captures the similarities among events in a longer temporal range, which potentially increase the connectivity of the temporal graph, while a small \( r \) only considers closely adjacent symbols as similar. In the extreme case when \( r \) approaches infinity, \( W^n \) becomes an almost constant matrix, since all appearing event pairs will be fully and equally connected.

The right part of Figure 1 provides a graphical illustration of the event sequence on the left part. In the sequence, we have 5 observations of 4 unique events. We show the duration between pairwise events. In this example, we use \( \Delta = 3 \) months and \( r = 5 \) days.

In our empirical study on real-world EHR data warehouse, we optimize \( r \) based on the algorithm performance in specific applications.

### 3.2 Temporal Signature Identification

With all the constructed temporal graphs, we want to identify the temporal signatures that can be used to best explain the observations. Our idea is to compute the graph bases as the temporal signatures which can be used to reconstruct the observed temporal graphs. In Figure 2, we have one simplified example, where we have three graph bases, and one observed graph can be expressed as the average of the first two bases. In practice, we do not know the bases at the beginning, and our temporal signature identification problem is exactly the process identifying the unknown graph bases with the observed temporal graphs.

We call the resultant graph bases as temporal phenotypes, which capture evolving patterns of the health conditions hidden in the event sequences. To be specific, suppose we have constructed the temporal graph \( G^n \) for each sequence \( s_n \), and \( G^n \) is associated with the adjacency weight matrix \( W^n \in \mathbb{R}^{M \times M} \). To reconstruct \( G^n \), we assume there are \( K \) graph bases \( B^k \in \mathbb{R}^{M \times M} \) for \( k = 1, 2, \cdots, K \), which can be used to approximate the adjacency matrix \( W^n \):

\[
W^n = \sum_{k=1}^{K} A_{nk} B^k,
\]

where \( A \in \mathbb{R}^{N \times K} \) is the matrix of reconstruction coefficients. To compute the optimal graph bases and the reconstruction coefficients, we minimize the total reconstruction error:

\[
\mathcal{J}(A, B) = \frac{1}{2} \sum_{n=1}^{N} \| W^n - \sum_{k=1}^{K} A_{nk} B^k \|_F^2,
\] (3)

where \( \| \cdot \|_F \) is the matrix Frobenius norm.

To make the solutions more interpretable, we also consider two constraints on the reconstruction coefficients in \( A \) and the graph bases \( B^k \) for \( k = 1, 2, \cdots, K \). The first constraint is about the non-negativity, i.e., \( B^k \succeq 0 \) for all \( k \),
since our original temporal graphs are non-negative. The second constraint requires $A \geq 0$ and $\sum_k A_{nk} = 1$, for $n = 1, \cdots, N$, which make the rows of $A$ to be valid multinomial distribution. In this way, we can quantify each patient by the temporal signatures with probabilities which can be in turn used for personalized medicine, patient segmentation, and disease diagnosis.

### 3.3 Regularization

As we introduced earlier, the reconstruction coefficients in $A$ can be used for various applications. In particular, for the medical diagnosis application, our goal is to derive informative features to improve the diagnosis performance, i.e., the classification of control/case groups for the patients. To this end, we extend the temporal signature identification for temporal graphs with regularization $\Omega(A) \geq 0$:

$$J(A, B) = \frac{1}{2} \sum_{n=1}^{N} \| W^n - \sum_{k=1}^{K} A_{nk} B^k \|_F^2 + \lambda \Omega(A),$$

where $\lambda \geq 0$ is the parameter controlling the degree of regularization. In the following, we propose several regularizations as $\Omega(A)$ to incorporate additional knowledge on the patients under study.

#### 3.3.1 Similarity based regularization

In the first case, we may have implicit similarity links between patients indicating they are from the same group (case or control). We can encourage the linked patients to have similar signatures in $A$ using the following regularization:

$$\Omega(A) = \frac{1}{2} \sum_{n_1, n_2} \frac{1}{2} \| A_{n_1} - A_{n_2} \|^2 S_{n_1 n_2},$$

where $S \in \mathbb{R}^{N \times N} > 0$ is symmetric matrix encoding the similarity information. Note that, when $S$ is asymmetric, we can just equivalently replace $S$ with $(S + S^T)/2$ without changing $\Omega(A)$. It follows that $S_{n_1 n_2} = S_{n_2 n_1}$ and

$$\Omega(A) = \frac{1}{2} \text{tr}(A' L A),$$

where $L = D - S$ and $D$ is the diagonal degree matrix such that $D_{nn} = \sum_{n'} S_{nn'}$. Note that, some rows/columns of $S$ may be completely zero if we do not have knowledge about the corresponding patients, e.g., the instances in the test set.

#### 3.3.2 Model based regularization

In the second case, we may have access to the group information of the patients. We let $Y_n = 1$ if the $n$-th patient is from the case group and $Y_n = -1$ if the patient is from the control group. With the explicit label information, we can define the regularization $\Omega(A)$ directly with a discriminative model $Pr(A_n, Y_n|H)$:

$$\Omega(A) = - \frac{1}{|\mathcal{L}|} \sum_{n \in \mathcal{L}} \log Pr(A_n, Y_n|H),$$

which is termed as average log-loss in the literature. Here, $\mathcal{L}$ is the training set where we have label $Y_n$ for $n \in \mathcal{L}$.

One particular choice for the discriminative model we can use for the case/control classification of patients is the logistic regression:

$$Pr(A_n, Y_n|H) = \frac{1}{1 + \exp(-Y_n f(A_n))},$$

where the linear model $H : A_n \mapsto f(A_n) = A_n \Theta + \theta$ and $(\Theta, \theta)$ are parameters in the model $H$. It follows that

$$\log Pr(A_n, Y_n|H) = - \log(1 + \exp(-Y_n f(A_n))).$$

1273
In addition to the log-loss for probabilistic model, other loss terms can also be used with the linear model $H$. We consider the hinge loss for $(A_n, Y_n)$:

$$\text{loss}(A_n, Y_n | H) = \max\{0, 1 - Y_n f(A_n)\},$$

and

$$\Omega(A) = \frac{1}{|C|} \sum_{n \in C} \text{loss}(A_n, Y_n | H). \quad (7)$$

The details on how to solve the optimization problem can be referred to [14].

4 Empirical Evaluation

In this section, we evaluate the effectiveness of our temporal signature identification approach on real-world patient Electronic Health Record (EHR) data warehouse, which includes the records of 319,650 patients over 4 years. We use the diagnosis information of the first three digits of ICD-9 (which stands for International Classification of Disease, 9th Version) code and the medication information in terms of drug ingredients to construct the EHR sequences. The temporal graphs are constructed from those sequences according to Definition 1. We will study the following two specific problems:

Hospitalization Prediction. We identify a set of 430 Congestive Heart Failure (CHF) patients with Chronic Obstructive Pulmonary Disease (COPD) pre-condition. Among them 100 are hospitalized within one year after CHF confirmation, the rest 330 patients are not. The goal is to make use of the records 360 days prior to the CHF confirmation date to predict whether the patients will be hospitalized or not within one year after CHF confirmation. The graphical illustration of this setting is in Figure 3 (a).

Disease Early Prediction. We first identify a set of 1127 case patients who are confirmed with Congestive Heart Failure (CHF), and then construct a set of 3850 group matched controls. For every patient, we set an operation criterion date, which is the CHF confirmation date for case patients, the day of the last record in our database for control patients. We then trace back from the operation criterion date, hold off the records within the prediction window (180 days), and use the records in observation window (360 days) for analysis. The graphical illustration of such setting is in Figure 3 (b).
use thicker edges to denote stronger weights (which suggests shorter intervals). We tested the different strategies introduced in Section 3 to learn the temporal phenotypes. The composition coefficients for every patient will be used as their vector representations for the prediction tasks. For comparison purpose, we also implemented the following baselines:

- **Aggregated Vector Representation (AV).** This method represents every patient as a vector with dimensionality equal to the number of distinct medical events in all patients’ EHR records. The value on a specific dimension is just the frequency of the corresponding event in his/her EHR sequence.

- **Bag-of-Pattern Representation (BP).** This method first run a standard sequential pattern mining algorithm to detect frequent patterns from the patients’ EHR sequences, and those patterns will be combined together to form a pattern repository. Every patient will be represented as a vector with dimensionality equal to the size of such pattern repository. The value on a specific dimension will be the frequency that pattern appeared in the EHR sequence of the corresponding patient.

- **Event Signature Representation (ES).** This method implements the temporal signature mining algorithm proposed in [22], which identifies the temporal patterns in patient EHRs via a constrained optimization procedure. The patients will still be represented by the bag-of-pattern representation as in the BPS method.

After the vector based representation for every patient is derived, we then adopt Support Vector Machine (SVM) to perform prediction. The classification performance is measured by Area Under the Receiver Operating Characteristics Curve (AUC), Area Under the Precision Recall Curve (AUPR), and Accuracy (ACC), and these measures are averaged over 10-fold cross validation. The parameters in our methods are tuned with a greedy grid search strategy as follows.

We first construct the temporal graphs using the unsupervised method, and the locality controlling factor $r$ is tuned with cross validation on the prediction results using the signatures detected from the constructed graph. Then the number of bases is tuned based on the prediction results with unsupervised signature identification. Finally the tradeoff parameter for regularized signature identification methods is tuned with $r$ and number of signatures fixed. In both studies, we set $\Delta = 90$ days, as both are chronic disease scenarios.

| Table 1 summaries the quantitative results with parameters chosen in the ways described above. From the table we can observe that: (1) Prediction using the signatures mined from our proposed methods can achieve better prediction performance compared to those baselines, which suggests the effectiveness of the proposed graphical scheme; (2) Regularized signature identification can produce better results because they utilizes supervision information.

In addition to the quantitative results, it would also be interesting to examine the validity of the detected temporal signatures qualitatively. We present three signatures with largest magnitude of the average composition coefficients for each data set in Figure 5 and 6 respectively. For all three signatures in Figure 5 we can clearly observe the drug hubs...
Temporal Signature Identification

Despite the promising quantitative and qualitative results, there are several aspects of this framework that we want to specifically discuss here.

- Comparing to sequence based representation, graph based representation is more compact, but at the same time it will suffer from information loss. For example, if a pair of consecutive events with the same temporal ordering repeat multiple times within the EHR of a specific patient, they will only appear once in the constructed temporal graph and the weight of the edge linking them will be aggregated from all the durations between their appearances. Usually the trends of the variations of such durations between repeated pairwise events can be indicative of the progression of some impending disease conditions. Therefore a smart way of encoding such
information to the constructed temporal graph would be a great enhancement to the current representation.

- The graph construction process would be computationally expensive if the number of distinct events is large, and the following basis learning process will be time consuming as well because it is correlated with the dimensionality of the graph adjacency matrix, which is the number of the distinct events. One potential strategy that could make this process more efficient is the multilevel/hierarchical graph learning approach [23], where we can construct a hierarchy of graphs from coarse to fine by utilizing the event ontology. The patterns can be detected from the coarsened graph efficiently and refined to more detailed levels.

- As can be seen from Figure 5 and 6, the detected temporal signatures could be messy because of the high variability of the patient EHRs. Besides the multilevel strategy, it would also be interesting to investigate sparsity induced regularizers, which can make the learned signatures more interpretable and clear.

6 Conclusion

In this paper, we proposed a novel graph based representation for patient EHRs, which encodes distinct medical events as well as their temporal relationships. Compared to traditional sequence and matrix based representations, graphs are more compact and intuitive. We presented several approaches to identify interesting temporal phenotypes based on such graph based representation, and validated their effectiveness on real-world data sets.

References


DenguePredict: An Integrated Drug Repositioning Approach towards Drug Discovery for Dengue

QuanQiu Wang¹, MS, Rong Xu², PhD,

¹ThinTek, LLC, Palo Alto, CA
²Department of Epidemiology and Biostatistics, Institute of Computational Biology, School of Medicine, Case Western Reserve University, Cleveland, OH

Abstract

Dengue is a viral disease of expanding global incidence without cures. Here we present a drug repositioning system (DenguePredict) leveraging upon a unique drug treatment database and vast amounts of disease- and drug-related data. We first constructed a large-scale genetic disease network with enriched dengue genetics data curated from biomedical literature. We applied a network-based ranking algorithm to find dengue-related diseases from the disease network. We then developed a novel algorithm to prioritize FDA-approved drugs from dengue-related diseases to treat dengue. When tested in a de-novo validation setting, DenguePredict found the only two drugs tested in clinical trials for treating dengue and ranked them highly: chloroquine ranked at top 0.96% and ivermectin at top 22.75%. We showed that drugs targeting immune systems and arachidonic acid metabolism-related apoptotic pathways might represent innovative drugs to treat dengue. In summary, DenguePredict, by combining comprehensive disease- and drug-related data and novel algorithms, may greatly facilitate drug discovery for dengue.

Introduction

Dengue is the most common vector-born viral infection in humans and the most rapidly spreading viral disease globally. Over 40% of the world’s population live in dengue-endemic areas, and about 50 to 100 million people are infected with the dengue virus every year. Currently, there are no curative drugs for dengue [1-3]. Therefore, cost-effective approaches are needed to rapidly discover innovative drug treatments for it. Drug repositioning is a drug discovery strategy that seeks to renew failed drugs or expand indications for approved drugs [4]. Currently, computational drug repositioning has not yet been applied to the search for drug treatments for dengue [5].

Disease genetics provide strong evidence to connect genes to human diseases. Variations in several genes have been shown to influence susceptibility and resistance to the dengue virus, as well as disease progression and severity [6-9]. These genes are involved in multiple genetic pathways associated with dengue as well as many other diseases. We hypothesize that diseases that share high genetic relevance with dengue may offer insights into disease biological basis and provide unique opportunities in developing effective drug treatments for dengue. Here we present a drug repositioning system (DenguePredict) that first finds diseases that are genetically related to dengue and then use dengue-related diseases as a window into understanding the biology of dengue and discovering drug candidates to treat it. Our study is different from current disease genetics-based drug discovery studies, which often directly infer drug targets from disease-associated genes [10-11]. To directly translate disease genetics into therapeutics, we need to know that disease-associated genes are involved in disease pathogenesis. However, the genetic basis of many diseases, including dengue, still remains unknown and the effect size of many disease-associated genes, for instance disease-associated genes discovered through genome-wide association studies (GWAS), is generally modest. Here we present an alternative strategy to circumvent these obstacles. We use disease genetics data as merely a starting point to infer interconnections among thousands of diseases and then develop a novel drug repositioning strategy to infer drug treatments based on these genetically related diseases and their associated drug treatments. Our intuition is that if two diseases share high genetic relevance, it is likely that these two diseases are related in pathophysiology even though the exact biology may remain unknown, therefore drugs that are effective in treating one disease may treat the other.

DenguePredict is a computation-based drug repositioning system. Computational drug repositioning approaches can be classified as drug-based, disease-based, and both [12-14]. Drug-based approaches leverage upon known drug molecular structures or functions such as chemical structure and properties, molecular docking, gene expression and drug side effects [15-21]. It was recognized that drug screens based on existing drugs might fail to identify new therapeutic mechanisms [22]. On the other hand, disease-based approaches put less emphasis on existing drugs and focus more on disease mechanisms and interrelationships, therefore have potential in discovering truly innovative
drugs. Disease-based approaches used disease-related data ranging from genome [10-11, 19-20] to phenome [23-27]. Many drug repositioning systems used well-established computational and statistical algorithms, including regression/classification, machine learning, network analysis, and text mining [14]. The keys to the success of a computational drug repositioning system include both the unique datasets included in the system as well as innovative ways in integrating various disease- and drug-related data towards specific problems (i.e. specific diseases or drugs).

There are three key components in DenguePredict. First DenguePredict contains a comprehensive drug-disease treatment relationship knowledge base (TreatKB) that we recently constructed from multiple heterogeneous and complementary data resources using advanced computational techniques including natural language processing, text mining and data mining [28-30]. TreatKB includes 9,216 drug-disease treatment pairs extracted from FDA drug labels, 111,862 pairs extracted from the FDA Adverse Event Reporting System (FAERS), a database supporting the FDA’s post-marketing drug safety surveillance, 34,306 pairs extracted from 22 million published biomedical literature abstracts, and 69,724 pairs extracted from 171,805 clinical trials. All together, TreatKB contains 208,330 drug-disease treatment pairs for 2484 drugs and 24,511 diseases. Second, we used disease genetics data from both the Online Mendelian Inheritance in Man (OMIM), a comprehensive database of human genes and genetic phenotypes [31], and the Catalog of Published Genome-Wide Association Studies from the US National Human Genome Research Institute (NHGRI), an exhaustive source containing the description of disease-trait-associated single nucleotide polymorphisms (SNPs) from published GWAS data [32]. We then enriched these disease genetics data by manually curating dengue-associated genes from published biomedical literature. Third, we used a novel signal prioritization algorithm that we recently developed [25] to find candidate drugs from dengue-related diseases.

Materials and methods

DenguePredict is depicted in Fig. 1 and consists of the following steps: (1) we constructed an integrated genetic disease network (GDN) using disease-gene associations and protein-protein interaction data from multiple large data resources. We applied a network-based ranking algorithm to find dengue-related diseases from GDN; (2) we examined what kinds of diseases and genetic pathways were enriched among top-ranked diseases; (3) we developed a drug repositioning approach to systematically transfer drugs from dengue-related diseases to treat dengue itself. We evaluated DenguePredict using the only two drugs that have been tested in clinical trials for the treatment of dengue; and (4) in order to better understand the top-ranked repositioned drug candidates, we determined which classes of drugs were enriched and which common genetic pathways these drug candidates target.

Fig.1. The drug repositioning and analysis pipeline of DenguePredict.

1. Construct an integrated genetic disease network (GDN) and find dengue-related diseases from GDN

1.1 Construct GDNs

We used disease-gene association data from three data resources to construct GDN. The first resource is the OMIM database [31]. We downloaded the OMIM database and mapped gene names to their corresponding approved human gene symbols as defined by the HUGO Gene Nomenclature Committee (HGNC) [33]. We extracted a total of 15,462 disease-gene pairs, representing 5,983 diseases and 8,831 genes.
The second source is the GWAS Catalog [32]. We mapped SNPs to their associated strongest genes, which were subsequently mapped to their corresponding approved human gene symbols as defined by the HGNC. In total, we obtained 22,470 disease/traits-gene pairs, representing 881 diseases/traits and 8,689 genes.

The third source is the published biomedical literature. We manually curated dengue-related biomedical literature and enriched dengue-associated genes included in OMIM and the GWAS Catalog. We classified curated dengue-gene pairs into genetics-based (to enrich data in OMIM) and genomics-based (to enrich data in the GWAS Catalog). Due to the intensive manual effort, we only curated dengue-associated genes from literature.

We first built two sub-networks separately using disease-gene associations from OMIM (GDN_OMIM) and the GWAS catalog (GDN_GWAS). We then integrated them into one network. On both sub-networks, two diseases were connected if their associated genes (proteins) interacted. The edge weights were determined by the numbers of protein-protein interaction (PPI) pairs between two diseases. The PPI data was obtained from the STRING database [34]. From the STRING database, we obtained a total of 4,137,054 human PPI pairs representing 17,756 human proteins. In building the integrated GDN, we mapped nodes in GDN_OMIM to the nodes in GDN_GWAS if the nodes represented the same diseases. The mapping was done based through the unified medical language system (UMLS) Concept Unique IDs (CUIs) [35]. The fact that only 29 diseases mapped between OMIM and GWAS Catalog demonstrate that the diseases in these two databases are largely complementary and that our mapping algorithm needs further improvements. For comparison, we also generated ten random networks by randomly shuffling the edges of the real GDN while maintaining the proportion of edges between diseases. The summary statistics of GDN and the two sub-networks are shown in Table 1.

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes (diseases) (n)</th>
<th>Edges (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDN_OMIM</td>
<td>4,848</td>
<td>882,751</td>
</tr>
<tr>
<td>GDN_GWAS</td>
<td>856</td>
<td>200,758</td>
</tr>
<tr>
<td>GDN</td>
<td>5,675</td>
<td>1,083,538, including 29 inter-network edges</td>
</tr>
</tbody>
</table>

Table 1. Numbers of nodes and edges for two sub-networks and the integrated GDN.

1.2 Find diseases that share high genetics with dengue from GDN

Recently we develop network-based ranking algorithms to prioritize genes for a given disease [27, 36], to prioritize diseases for a given disease [25], and to prioritize diseases for a given microbial metabolite [37]. In this study, we applied these network-based ranking algorithms to find diseases that share high genetics with dengue. The iterative ranking algorithm is defined as: $p^{t+1} = (1-r)Wp^t + rp^0$, wherein $W$ is the column-normalized adjacency matrix of the integrated GDN (the equal transition probability between GDN_OMIM and GDN_GWAS) and $p^0$ is a vector in which the $i$-th element held the normalized ranking score of disease $i$ at $t$-th iteration. The initial probability vector $p^0$ contains dengue with a probability of 1.0. Other diseases are then ranked according to the steady-state probability vector, which is obtained by iterating the algorithm until the change between $p^{t+1}$ and $p^t$ is less than $10^{-6}$.

2 Analyze top-ranked diseases that share a high degree of genetic similarity with dengue

2.1 Analyze disease classes among ranked dengue-related diseases

To systematically understand dengue-related diseases, we examined disease classes enriched among top-ranked diseases retrieved from GDN with dengue as the input. We classified these diseases into sixteen categories using the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD10) [38]. The ICD10 includes 22 highest-level disease classes. We used 16 of the 22 chapters and excluded six non-specific disease classes such as “Codes for special purposes,” and “Injury, poisoning and certain other consequences of external causes.” Since the terms used in ICD10 are often different from those in GDN, we mapped disease terms in ICD10 to their synonyms through UMLS CUIs [35]. We retrieved a list of ranked dengue-related diseases from GDN. For diseases at 10 different ranking cutoffs (top 10%, 20%, ... 100%), we calculated percentages of the sixteen disease classes among them.

2.2 Analyze genetic pathways shared among ranked dengue-related diseases

To gain insights into common mechanistic relationships shared among dengue-related diseases, we analyzed genetic pathways associated with them. Functions of highly enriched pathways might provide insights into common molecular mechanisms shared among dengue-related diseases. First, we retrieved disease-associated genes from OMIM and from the GWAS catalog. We ranked each gene based on how many of the dengue-related diseases it was
associated with as well as the ranking scores of those diseases: 
\[ R_{\text{gene}} = \sum_{i=1}^{n} R_{\text{disease},i} \],
where \( n \) is the number of dengue-related diseases that the gene is associated with and \( R_{\text{disease},i} \) is the disease ranking score (as determined by the network-based disease ranking algorithm described above). We then analyzed gene-associated pathways using pathway data from the Molecular Signatures Database (MSigDB), a collection of 10,295 annotated genetic pathways or gene sets from multiple sources [39]. We ranked these pathways based on the number of genes associated with dengue-related diseases as well as the rank scores of those genes: 
\[ R_{\text{pathway}} = \sum_{i=1}^{n} R_{\text{gene},i} \],
where \( n \) is the number of genes that the pathway contains and \( R_{\text{gene},i} \) is the gene ranking score as determined above. We compared top-ranked pathways for GDN to those for random GDNs in order to identify pathways enriched for dengue-related diseases.

3 Reposition drugs associated with dengue-related diseases to treat dengue

3.1 Drug repositioning algorithm
We ranked FDA-approved drugs by the number of dengue-related diseases that they could treat as well as the ranking scores of those diseases. The drug prioritization algorithm is defined as: 
\[ R_{\text{drug}} = \sum_{i=1}^{n} R_{\text{disease},i} \],
wherein \( n \) is the number of dengue-related diseases that can be treated by a drug and \( R_{\text{disease},i} \) is the disease ranking score (as determined by the network-based disease ranking algorithm described above).

3.2 De-novo validation using dengue drugs tested in clinical trials
The inputs to the drug prioritization algorithm are a ranked list of dengue-related diseases (output from the disease ranking algorithm) and their associated drug treatments as determined by the drug-disease treatment pairs from four TreatKBs. Since the inputs contain neither dengue nor dengue-related treatment information, thereby our evaluation is in fact a de novo validation approach. Since no FDA-approved drugs are currently available for the treatment of dengue, we used drugs that have been tested in clinical trials for evaluation. We retrieved a total of 101 dengue-related clinical trials from ClinicalTrials.gov (www.clinicaltrials.gov). While most of these trials test vaccines, five trials tested five different drugs: chloroquine (NCT00849602), ivermectin (NCT020445069), balapiravir (NCT01096576), celgosivir (NCT01619969), and uv-4B (NCT020661358). Among these five drugs, only chloroquine and ivermectin are FDA-approved drugs: chloroquine approved for treating malaria and amebiasis, and ivermectin approved for treating onchocerciasis and strongyloidiasis. For our drug repositioning purpose, we used the two FDA-approved drugs (chloroquine and ivermectin) as the gold standard. We calculated the rankings of these two drugs among all FDA-approved drugs. The higher these two gold standard drugs were ranked, the better the ranking algorithm was. We compared the rankings of these two drugs derived from GDN to those from ten randomly generated GDNs. In addition, we also compared the performances across the four different TreatKBs.

4 Analyze top-ranked drug candidates

4.1 Analyze drug classes among ranked drug candidates
We examined which classes of drugs were enriched among top-ranked repositioned drug candidates. We classified drugs using drug classes defined by the Anatomical Therapeutic Chemical (ATC) classification system [40]. The ATC system consists of 13 first-level codes, which were further classified into 94 second-level codes, 267 third-level codes, 882 fourth-level codes, and 4,580 fifth-level codes, which are individual drugs. We experimented classifying drugs using different level ATC codes and found that the third level ATC codes provided sufficient but not too fine-grained granularity.

4.2 Analyze genetic pathways targeted by repositioned drug candidates
To understand the common mechanisms of action underlying top-ranked repositioned drug candidates, we analyzed genetic pathways targeted by these drug candidates. The method for drug pathway analysis was similar to that used for disease pathway analysis as described above, except that the drug-target gene association data was from DrugBank [41]. We compared top-ranked pathways based on GDN to those based on random GDNs in order to identify pathways enriched for repositioned dengue drug candidates. Functions of these enriched pathways might provide insights into common molecular mechanisms targeted by drug candidates. We performed literature search for supporting evidence that these enriched pathways might be targeted for dengue treatments.
Results

1. Infectious and parasitic diseases, neoplasms, and diseases of the digestive system are enriched among top-ranked dengue-related diseases

Using dengue as the input, we retrieved a ranked list of 4729 diseases from GDN. The disease class “Certain infectious and parasitic diseases” was highly enriched among top-ranked diseases: 9.73% among top 10% ranked diseases as compare to 1.46% among all diseases. This is expected, since dengue is known to share genetics with other infectious diseases such as malaria, Mycobacterium tuberculosis, and HIV [42-44]. Therefore, the enrichment of this disease class among diseases retrieved from GDN roughly served as a positive control in validating both network construction and disease ranking algorithms. Interestingly, two other disease classes “Neoplasms” and “Diseases of the digestive system” were also significantly enriched among top-ranked diseases. For comparison, none of the sixteen disease classes were enriched when the randomly generated GDNs were used (data not shown).

Fig. 2. Three ICD10AM disease classes were enriched among top-ranked diseases retrieved from GDN.

2. Immune-related pathways may be the common mechanisms underlying dengue and its related diseases

Fig. 3 shows the top fifteen enriched pathways and their percentages among all retrieved pathways. Among these enriched pathways, at least nine pathways are involved in human immune system, including “REACTOME IMMUNE SYSTEM,” “KEGG FC EPSILON RI SIGNALING PATHWAY,” and “REACTOME ADAPTIVE IMMUNE SYSTEM.” Experimental and observational findings in past years suggest that immune system-related mechanisms are involved in dengue pathophysiology. Dengue fever is characterized by thrombocytopenia and vascular leak with altered plasma cytokine profiles. Suggested immune mechanisms include platelet activation and apoptosis modulating monocyte inflammatory responses, increased levels of mediators like tumor necrosis factor-α and interleukin-1β [45], interplay between plasmablasts, platelets, and complements [45], disruption of the interaction of Daxx and NF-kB to induce CD137-mediated apoptosis [47], immunodominance changes [48], the involvement of Notch signaling pathways to modulate host adaptive immune response and altered profiles of cytokines produced by cross-reactive T cells [49].
Four cancer-related pathways were also enriched, including “KEGG PATHWAYS IN CANCER,” “KEGG ENDOMETRIAL CANCER,” “KEGG PROSTATE CANCER,” and “KEGG COLORECTAL CANCER.” The same enrichment for cancers was observed in above disease class enrichment analysis. However, we could not find literature evidence supporting the direct relationship between dengue and cancers. Since immunopathogeneseis are known to be involved in many cancers and dengue, common immune mechanisms may underlie both dengue and cancers. Studies have shown that carica papaya leaves exhibits both anti-tumor activity and immunomodulatory effects in dengue [50-51]. In summary, by systematically analyzing genetic pathways involved with dengue-related diseases, we can gain deeper insights into molecular mechanisms underlying dengue and its related diseases.

3. DenguePredict found the two clinical trial dengue drugs and ranked them highly
As shown in Table 2, DenguePredict consistently ranked chloroquine highly (ranging from top 0.96% to 5.98%). Ivermectin was ranked lower than chloroquine, but still significantly higher than those derived with random GDNs. Comparing across four TreatKBs, DenguePredict using the MEDLINE-based TreatKB performed the best, with chloroquine ranked at 0.96% and ivermectine at 22.75%. Combining drug-disease treatment pairs from all four databases did not improve the performance (data not shown). The fact that we ranked both chloroquine and ivermectin highly demonstrated the validity of our repositioning strategy. The output of DenguePredict is a list of FDA-approved drugs ranked based on their likelihood for treating dengue.

<table>
<thead>
<tr>
<th>TreatKB</th>
<th>GDN</th>
<th>GDN_Random</th>
<th>Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chl</td>
<td>Iver</td>
<td>Chl</td>
</tr>
<tr>
<td>FDA-approved</td>
<td>4.65%</td>
<td>54.85%</td>
<td>20.85%</td>
</tr>
<tr>
<td>Post-market</td>
<td>5.98%</td>
<td>43.04%</td>
<td>24.69%</td>
</tr>
<tr>
<td>ClinicalTrials</td>
<td>5.83%</td>
<td>56.76%</td>
<td>18.56%</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>0.96%</td>
<td>22.75%</td>
<td>8.37%</td>
</tr>
</tbody>
</table>

Table 2. Drug reposition evaluation across four different TreatKBs.

4. Immune system-related drugs are highly enriched among top-ranked repositioned drug candidate
The top 10 ranked repositioned drug candidates using the four TreatKBs are shown in Table 3. Drugs that were consistently ranked highly across four TreatKBs include many corticosteroids (i.e. methylprednisolone, dexamethasone, prednisone, and prednisolone).

<table>
<thead>
<tr>
<th>Rank</th>
<th>FDA-approved</th>
<th>Post-market</th>
<th>ClinicalTrials</th>
<th>MEDLINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Methylprednisolone</td>
<td>Dexamethasone</td>
<td>Chlordiazepoxide</td>
<td>Nitric oxide</td>
</tr>
<tr>
<td>2</td>
<td>Betamethasone</td>
<td>Prednisolone</td>
<td>Bevacizumab</td>
<td>Heparin</td>
</tr>
<tr>
<td>3</td>
<td>Triamcinolone</td>
<td>Nitric oxide</td>
<td>Cisplatin</td>
<td>Methotrexate</td>
</tr>
<tr>
<td>4</td>
<td>Dexamethasone</td>
<td>Prednisone</td>
<td>Paclitaxel</td>
<td>Celecoxib</td>
</tr>
<tr>
<td>5</td>
<td>Prednisolone</td>
<td>Fentanyl</td>
<td>Carboplatin</td>
<td>Prednisolone</td>
</tr>
<tr>
<td>6</td>
<td>Prednisone</td>
<td>Methylprednisolone</td>
<td>Gemcitabine</td>
<td>Iron</td>
</tr>
<tr>
<td>7</td>
<td>Cortisone</td>
<td>Aspirin</td>
<td>Doxorubicin</td>
<td>Adenosine</td>
</tr>
<tr>
<td>8</td>
<td>Hydrocortisone</td>
<td>Methotrexate</td>
<td>Dexamethasone</td>
<td>Vitamin c</td>
</tr>
<tr>
<td>9</td>
<td>Fluorouracil, 5-fu</td>
<td>Acetaminophen</td>
<td>Cyclophosphamide</td>
<td>Indomethacin</td>
</tr>
<tr>
<td>10</td>
<td>Mometasone</td>
<td>Celecoxib</td>
<td>Prednisone</td>
<td>Resveratrol</td>
</tr>
</tbody>
</table>

Table 3. Top 10 drug candidates repositioned from four TreatKBs.

We examined which classes of drugs were enriched and analyzed genetic pathways associated with highly-ranked drug candidates. As shown in Fig. 4, the most enriched (more than 25-fold enrichment as compared to random GDNs) drug class was “Other analgesics and antipyretics.” This finding is consistent with the current symptomatic and supportive treatment for dengue using analgesic-antipyretic therapy for the relief of lethargy, malaise, and fever associated with the disease [2]. The second most enriched drug class was “anti-inflammatory and anti-rheumatic products” (15-fold enrichment), which includes corticosteroids. Corticosteroids are potent anti-inflammatory agents with a wide range of effects on the immune system. Observational studies have suggested that corticosteroids may benefit people with dengue-related shock and may prevent disease progression [52]. Our study provides independent mechanistic evidence supporting treatment benefits of corticosteroids in the treatment of dengue.
5. Repositioned drug candidates mainly target apoptosis- and immune-related pathways

Fig. 5 shows top 21 enriched pathways for repositioned drug candidates. The pathway involved in arachidonic acid metabolism showed 28-fold enrichment as compared to random networks. Arachidonic acid (AA) is a lipid second messenger generated by hydrolysis of membrane phospholipids via phospholipase A2 (PLA2). Malewicz et al. reported that dengue virus was able to activate PLA2 and generate AA [53]. Jan et al. showed that AA, superoxide anion, and NF-kappa B are sequentially involved in dengue virus-triggered apoptotic pathways in human neuroblastoma cells and that inhibition of PLA(2) activity by the PLA(2) inhibitors diminished DEN-2 virus-induced apoptosis [54]. Many of these top enriched pathways are related to immune systems, including natural killer cell mediated cytotoxicity, complement pathway, immunoregulatory interactions, and B lymphocyte pathways. This is consistent with the pathway analysis for dengue-related diseases. In summary, our study indicates that repositioned drug candidates that target arachidonic acid metabolism and/or immune systems might benefit people with dengue. While we did not perform literature search for all other enriched pathways, further investigating these pathways may generate novel hypotheses for dengue drug discovery.
Discussion
By leveraging upon vast amount of knowledge of disease genetics, drug targets, protein interactions, and drug treatments, DenguePredict effectively ranked the only two drugs currently under clinical trial for the treatment of dengue highly. Our approach also suggested potential genetic pathways involved in disease mechanisms and mechanisms of actions of the drug repositioning candidates, which warrants further investigation. DenguePredict is highly generalizable and can easily be retargeted to find drug candidates for other genetic diseases. We expect that its performance will vary among specific diseases and critically depends on the data (both available disease genetics and drug treatments) included for each disease.

We were unable to compare DenguePredict to existing computational drug repositioning systems since these systems did not include dengue in their study. For example, in a recent study, Gottlieb et al used disease-disease similarities and drug-drug similarities from multiple databases to construct a classifier (PREDICT) to determine treatment associations between 593 drugs and 313 diseases [24]. While PREDICT is among currently most comprehensive drug repositioning systems, dengue was not included. On the other hand, DenguePredict included significantly more drugs and diseases: 5,675 diseases on GDN; 2,484 drugs and 24,511 diseases in TreatKBs.

Our study can be further improved in several aspects. First, we can further reprioritize the generated ranked list of repurposed drug candidates by their costs. Dengue is most prevalent in developing countries; therefore costs of drugs needs to be considered. Second, our study generated lists of putative genetic pathways that might be involved with dengue-related diseases and drug candidates, however these candidates will evolve as new disease-associated genes are discovered. For example, as new genes are discovered for dengue or other diseases, more diseases will be linked to dengue from the updated genetic network. Third, we validated our repositioning algorithm using the two clinical trial dengue drugs. Due to the small size of our testing data, we are still uncertain about the precision of top-ranked drug candidates. Experimental or clinical studies are needed to test those candidates.

Competing interests
None.

Author’s contributions
RX and QW have jointly conceived the idea, designed and implemented the algorithms, and prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgements
RX was supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under the NIH Director’s New Innovator Award number DP2HD084068, the Training grant in Computational Genomic Epidemiology of Cancer (CoGEC) (R25 CA094186-06), and Case Western Reserve University/Cleveland Clinic CTSA Grant (UL1TR000439). The work done by QW was supported by above funding resources.

References


Mortality Prediction in ICUs Using
A Novel Time-Slicing Cox Regression Method

Yuan Wang, PhD¹, Wenlin Chen¹, Kevin Heard², Marin H. Kollef, MD²,
Thomas C. Bailey, MD², Zhicheng Cui¹, Yujie He¹, Chenyang Lu, PhD¹, Yixin Chen, PhD¹
¹Department of Computer Science and Engineering, Washington University, St. Louis, MO; ²School of Medicine, Washington University, St. Louis, MO

Abstract
Over the last few decades, machine learning and data mining have been increasingly used for clinical prediction in ICUs. However, there is still a huge gap in making full use of the time-series data generated from ICUs. Aiming at filling this gap, we propose a novel approach entitled Time Slicing Cox regression (TS-Cox), which extends the classical Cox regression into a classification method on multi-dimensional time-series. Unlike traditional classifiers such as logistic regression and support vector machines, our model not only incorporates the discriminative features derived from the time-series, but also naturally exploits the temporal orders of these features based on a Cox-like function. Empirical evaluation on MIMIC-II database demonstrates the efficacy of the TS-Cox model. Our TS-Cox model outperforms all other baseline models by a good margin in terms of AUC.PR, sensitivity and PPV, which indicates that TS-Cox may be a promising tool for mortality prediction in ICUs.

Introduction
Over the last few decades, machine learning and data mining techniques are increasingly used for ICU mortality prediction. For example, PhysioNet/CinC organized a challenge in 2012 to develop methods for patient-specific prediction of in-hospital mortality [1], where the data includes time series of vital signs during the 48 hours after ICU admission. There are two tracks in this competition concerning binary outcome measurement (dead or alive) evaluated by sensitivity and positive predictive value (PPV), and risk estimation evaluated by a range-normalized Hosmer-Lemeshow statistic. A great deal of models submitted by participants largely outperformed the baseline algorithm (SAPS-1). For example, Johnson et al [2] used a novel Bayesian ensemble learning algorithm, and Krajnak et al [3] combined machine learning and clinical rules to build a predictive model. Existing algorithms do not fully take advantage of the temporal changes of the input clinical time-series data. Models such as logistic regression (LR), support vector machine (SVM), and decision trees (DT) can only handle static and fixed-length feature vectors, which greatly limit their use on time-series. In this regard, it makes more sense to resort to Cox regression since it takes into account the temporal orderings of various features and it has widespread uses in survival analysis. In the following, we mainly survey the literature related to Cox regression in the area of medical care utilization.

Cox regression [4,5], though invented in the seventies, is still receiving significant interests from the domains of both healthcare and information technology. Different from methods mentioned above, Cox regression is one of the few models that take into account the temporal structure of events, which is very important for using time-series data to predict morbidity, readmission and mortality and to identify the associated factors. One of the classical examples for predicting morbidity is the risk assessment of cardiovascular disease [4]. A few studies also applied the Cox model for the prediction of outcome in patients with diabetes [6,7], stroke prediction [8], as well as ecologic studies [9]. These algorithms use the vast domain-specific knowledge that has been accumulated on the disease to manually select a limited number of risk factors and then put them into a Cox model.

Recent studies have also recommended using various machine-learning techniques combined with Cox regression for diverse topics in healthcare [10,11]. In addition, there have been a great effort towards making Cox regression more generalizable by adding regularizations to the parameters [12-15]. The above methods typically work on relational data with factors such as age, gender, race, weight, ecologic exposures, lab variables and clinical risk factors. Although some factors may be time-varying, they are typically processed in the canonical way. However, few have used Cox regression to classify multi-dimensional time-series
data, due to the challenge that Cox regression is originally proposed for survival analysis and cannot be directly used in time-series classification.

To address this problem, we propose in this paper a novel approach named Time Slicing Cox regression (TS-Cox), which not only takes advantage of Cox regression, but also can be applied on clinical time-series data. Specifically, the rich time-series data allow us to used sophisticated features such as statistical moments, powers in frequency domain, estimates in chaos theory and entropies in information theory as a novel set of covariates to predict mortality risk. Then, based on time-to-event characteristics of ICU data, we divide each time series into multiple windows for each waveform and extract features from every sliding window via signal analysis. Next, we employ Cox regression on those windows together with their associated time stamp information. After the model is trained, we are able to compute a hazard ratio as risk score for each waveform based on the extracted features and learnt model. Our TS-Cox model allows us to leverage on both powerful time-series features and the survival information in the time domain. To our knowledge, this is the first work that adapts Cox regression to classify multi-dimensional time-series.

The remainder of this paper is organized as follows. First, we introduce some background on the original Cox model and describe the proposed TS-Cox model. Next, empirical results of our TS-Cox model compared with baseline methods are presented and discussed. Finally, we summarize the research work and draw conclusions.

**Methods**

**Feature extraction method**

The input of our model is based on the features extracted from the multiple time-series of each patient. Many of our features are inspired by classic signal analysis methods including RR series distribution patterns (Mean, Median, Variance, Skewness, Kurtosis), magnitude of variability in the time domain (SD, RMSSD), and linear estimates in the frequency domain (power of VLF, LF, HF and LF/HF). A complete recipe for extracting these physiological features can be found in [16].

Besides, we also make use of features kindled by chaos theory and information theory, which are helpful in analyzing the degree of self-affinity and randomness of the time-series. Specifically, we adopt the detrended fluctuation analysis (DFA) based on chaos theory to estimate the statistical self-similarity of a signal [17-19]. We also compute the randomness measurements including approximate entropy [20,21], sample entropy [22,23] and permutation entropy [24]. They are measurements designed to quantify the degree of regularity versus unpredictability, reflecting the unpredictability of fluctuation in a time-series. A low value of the entropy indicates that the time-series is deterministic while a high value means that the time-series is unpredictable. These entropies are good indicators for cardiovascular signals where the occurrence of disease is highly correlated with the decrease of entropy. In summary, in this study we extracted 16 features from two types of time-series: heart rate (HR) and peripheral capillary oxygen saturation (SpO2), leading to 32 features in total for each time window of a patient.

**Classification method**

**Cox Regression**

Cox regression is a popular statistical model in survival analysis which deals with analysis of time duration until the happening of one or more events, such as sudden deterioration of a patient [4]. It estimates the risk of the occurrence of a particular event at a particular time stamp by quantifying the relation between the features and the risks.

Now we describe the Cox regression model. Let $X=(x_1, x_2, ..., x_m)$ be the feature vector for each patient where $m$ is the number of features. In this paper, the features come from the aforementioned signal analysis method. Let $h(t)$ be a hazard function assessing the instantaneous risk of an event (e.g. mortality) at time $t$. The Cox regression expresses the logarithm of the mortality risk $h(t)$ as a weighted sum of all features with a base function as follows:

$$\log h(t \mid X) = \alpha(t) + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m,$$

where $\beta_i$ is the coefficient of feature $x_i$, reflecting the relation between this feature and the risk. For example, a large $\beta_i$ means that the raise of the value of feature $x_i$ increases the risk of mortality. $\alpha(t)$ is a base function independent
from the feature vector $X$ whose output varies over time depending on $t$. Given the above equation (1), we have that

$$ h(t \mid X) = \frac{h_0(t)}{\prod_{i=1}^{m} \exp(\beta_i X_i)} $$

where $h_0(t)$ is called baseline hazard function and the following product is called proportional hazards. The proportional hazards assume that each feature contributes an independent multiplicative factor to the primary risk hazard rate. Cox regression estimates the parameters $\beta$ using maximum likelihood estimation over the partial log-likelihood function in which $h_0(t)$ is cancelled out. Consequently, we do not need to know the functional form of $h_0(t)$ as the output does not depend on the feature vector. Given a fixed time stamp $t$, the risk hazard rate $h(t, X)$ only depends on the feature values. Therefore, when computing the risk hazard rate at a particular time stamp in the future, we only need to compute the proportional hazards as a proxy for risk hazard rate.

The problem of survival analysis is further compounded by the presence of censored data [25,26], where the status of an instance is not observable after a certain time point. For example, an instance in mortality analysis is censored when the patient at question is still alive at the end of observation. Censored times $C_i$ are associated with each instance $i$ along with observed time for the event $O_i$. Define the failure time $T_i$ for instance $i$ as the minimum of $O_i$ and $C_i$, i.e., $T_i = \min(O_i, C_i)$. $O_i \leq C_i$ indicates that the event of interest has occurred within the censoring time. However, if $O_i$ is unknown then $T_i$ is set to $C_i$ and the instance is censored. Take hospital readmission prediction as an example. An event is defined as the onset of heart failure readmission within 30 days of discharge from the previous admission. The censored cases can be identified when (i) the patients whose follow-up details were lost over time or (ii) the patients was not readmitted within the time period of follow-up until the end of the study (which is fixed to 30 days in this case). This is commonly called the right censoring setting, which is the most frequently studied censoring phenomena in survival analysis.

**Time Slicing Cox regression (TS-Cox)**

Now we introduce our Time Slicing Cox regression (TS-Cox) method, which applies the Cox model to time-series data. Our method contains three steps: preprocessing, training, and deployment. We describe these three steps in detail as follows.

**Preprocessing.** We assume that the time-series has a basic unit step which is the time gap between two measurements. For example, in our data from MIMIC-II, each unit step is 60 seconds. For each patient, we maintain a sliding window from the beginning of each of his or her time-series. The window is a fixed-length time frame (e.g., window size =1024 unit steps) that can be regarded as a snapshot of the status of a patient. At each iteration, we extract features from the time window and record the time stamp of this window. We then move the window by a fixed step-size and repeat until we reach the end of the patient’s time-series.

For each time window, we adopt the aforementioned feature extraction method to obtain features from the time-series within the window. In our case, each time window is associated with two time-series for HR and SpO2. We extract 16 features from each time-series separately, leading to 32 features in total for each time window.

In each iteration, we move the window by a pre-specified step size (e.g., 100 unit steps) and extract features again for the new window. This way, we have multiple windows for each patient. Suppose the number of windows for a patient is $L$. Each window is also associated with a new failure time, which is the time gap between the end of the window and the original failure time $T$. Let $D_l$ be the last time stamp of window $l$, $l = 1, ..., L$, and $S_l$ be the failure time for window $l$, we have that $S_l = \max(T - D_l, 0)$. At the end of preprocessing, for each patient we have $L$ windows where each window $l$ is associated with 32 features and a failure time $S_l$ reflecting the new survival time for that window.

**Training.** The training procedure is to estimate the value of $\beta$, which largely follows the same workflow for Cox regression. Different from the original Cox regression where each training instance is a patient, our TS-Cox model counts each window of each patient as a training instance. In other words, each window is viewed as a unique patient. This way, the TS-Cox model considers the time-varying migration of features and the temporal ordering of all the time windows. In contrast, if we extract features directly from the entire time-series for all patients and then input them into models such as LR or SVM, we would lose such temporal information.
We still learn a weight vector $\beta$ using maximum likelihood estimation after training the TS-Cox model. For example, if there are 32 features for each time window, the length of the vector $\beta$ is 32. The training process is to tune the value of $\beta$ so that the resulting risk hazard rates for all training instances best respect the order of their failure time $S_t$.

**Deployment.** With the learnt $\beta$ in place after training, the evaluation procedure is to evaluate the risk of mortality for an out-of-sample patient. Specifically, given a patient, we extract the 32 features from their last time window using the same window size as in preprocessing and training. We then compute the proportional hazard $h'(X)$ which is the exponential of the dot product of $\beta$ and the feature vector $X$, i.e. $h'(X) = \exp(\beta X)$ as the hazard risk rate. To use the proportional hazard rate for binary classification, we need to set a threshold $c$ so that those patients with $h'(X)>c$ belong to one class and the rest belongs to another class. The threshold $c$ is typically chosen by aiming at a prescribed specificity, sensitivity, or positive predictive value (PPV) based on cross-validation on the training data.

In practice, our model can be deployed to monitor patients in real-time. We can move the time window to the last point as new data come in, extract updated features, and re-compute the hazard rate by feeding the updated features into the TS-Cox model. An alert will be triggered when the criterion (such as $h'(X)>c$) for certain event is met.

**Results and Discussion**

We conduct experiments to evaluate the performance of the proposed method. In our experiments, we used a publicly available ICU database, namely Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) [27]. The data harnessed in MIMIC-II were collected from the ICUs of Beth Israel Deaconess Medical Center from 2001 to 2008 and covers 26,870 adult hospital admissions (version 2.6). Two types of data were obtained: clinical data and physiological waveforms. The clinical data were acquired from the CareVue Clinical Information System (models M2331A and M1215A; Philips Healthcare, Andover, MA) and the hospital’s electronic archives. The data included patient demographics, nursing notes, discharge summaries, continuous intravenous drip medications, laboratory test results, and nurse-verified hourly vital signs, etc. The physiological waveforms were collected from bedside monitors (Component Monitoring System Intellivue MP-70; Philips Healthcare) and included high-resolution (125 Hz) waveforms (e.g., electrocardiograms), derived time series such as HR, blood pressures, SpO2, and monitor-generated alarms [28]. Here, we only extracted the time series of HR and SpO2 as the inputs to our model. Finally, our experiment datasets consist of 930 waveforms including complete HR and SpO2 records, among which there are 56 dead and 874 censored records. Our task is to predict whether a patient will die in the ICU.

We compare our method with the following baselines: 1) Classical classifiers including logistic regression (LR), linear SVM (SVM-I) and SVM with RBF kernel (SVM-r). Specially, we follow the traditional approach of first extracting features from the whole time series and then apply those classifiers for binary classification. 2) Time-Slicing version of logistic regression (TS-LR), linear SVM (TS-SVM-I) and SVM with RBF kernel (TS-SVM-r). In particular, we follow the same time-slicing mechanism as in TS-Cox and train these models on all time windows to learn the weight of each feature. When making prediction, we apply these learned weights on the last window of each waveform. For all time-slicing algorithms, we use 1024 unit steps as window size and 100 unit steps as step size for moving windows based on the previous experiments.

The dataset is highly imbalanced with the majority of patients being non-dead. Therefore, accuracy performance does not make the most sense in this case since one can achieve a high accuracy by always predicting non-dead for every patient. As a result, for evaluating the performance, we fixed the specificity of each model to about 0.95 and compare the sensitivity and PPV of each model. What’s more, we also evaluate the area under the ROC curve (AUC_ROC) and area under the precision-recall curve (AUC_PR). All the results reported in Table 1 are based on 5-fold cross validation. That is to say, we divide the whole dataset into 5 subsets where each subset contains 186 waveforms. Each algorithm is evaluated on each subset with all other 4 subsets being the training dataset. The final result is the mean of all evaluated subsets.

Table 1 below shows the results of all algorithms with specificity being fixed to 0.95. We make the following observations:
First of all, the proposed Cox regression with time slicing is better than the original Cox regression regardless of any measurements, which demonstrates the efficacy of using time slicing techniques in the Cox regression model. The advantage of Cox regression is that it takes time-to-event into consideration, while other models do not. Therefore, time slicing techniques work for Cox regression, but do not help other models since it merely increases the training instances for these classifiers.

Second, TS-Cox has both the largest AUC_ROC and the largest AUC_PR. Furthermore, TS-Cox surpasses all other models by a larger margin in regard of AUC_PR than of AUC_ROC, which signify that it makes more sense to measure AUC_PR than to estimate AUC_ROC for a skewed class distribution (most people still alive).

Third, the TS-Cox model constantly outperforms all the baseline models in terms of sensitivity and PPV, which indicate that TS-Cox is superior to other prediction models. We attribute this promising progress to the fact that the TS-Cox model is able to exploit the rich time-variant information of clinical time-series data.

Therefore, TS-Cox has a potential advantage over other methods to predict mortality in adult ICU monitoring system.

<table>
<thead>
<tr>
<th></th>
<th>AUC_ROC</th>
<th>AUC_PR</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.7463</td>
<td>0.1642</td>
<td>0.9485</td>
<td>0.1352</td>
<td>0.1455</td>
</tr>
<tr>
<td>SVM-l</td>
<td>0.4669</td>
<td>0.0763</td>
<td>0.9496</td>
<td>0.0643</td>
<td>0.0764</td>
</tr>
<tr>
<td>SVM-r</td>
<td>0.6823</td>
<td>0.1247</td>
<td>0.9485</td>
<td>0.1012</td>
<td>0.1100</td>
</tr>
<tr>
<td>TS-LR (Step=100)</td>
<td>0.7181</td>
<td>0.1826</td>
<td>0.9498</td>
<td>0.1652</td>
<td>0.2253</td>
</tr>
<tr>
<td>TS-SVM-l (Step=100)</td>
<td>0.5452</td>
<td>0.0938</td>
<td>0.9500</td>
<td>0.0456</td>
<td>0.0785</td>
</tr>
<tr>
<td>TS-SVM-r (Step=100)</td>
<td>0.6841</td>
<td>0.1402</td>
<td>0.9504</td>
<td>0.1336</td>
<td>0.1649</td>
</tr>
<tr>
<td>TS-Cox (Step=100)</td>
<td>0.7514</td>
<td>0.2522</td>
<td>0.9485</td>
<td>0.3424</td>
<td>0.2956</td>
</tr>
</tbody>
</table>

Conclusions
This paper proposed an integrated architecture for predicting death events in adult ICU patients based on a combination of time-series feature extraction methods and a novel time-slicing Cox regression method. The former produces more discerning features, and the latter exploits the strategy of slicing time windows on the static Cox model according to its time-to-event characteristics. We compared our TS-Cox model to conventional Cox model as well as LR, SVM-l, SVM-r, and their time-slicing version. The experiments indicate that TS-Cox has the highest AUC_PR, sensitivity and PPV by a huge margin at a specificity value around 0.95. These promising results demonstrate the feasibility of our new framework for better mortality prediction in adult ICU monitoring system.

Acknowledgements
This work was partially supported by CCF-1215302, IIS-1343896, and DBI-1356669 grants from the National Science Foundation of the USA, a Microsoft Research New Faculty Fellowship, a grant from the Barnes-Jewish Hospital Foundation, and a URSA grant from the Washington University.
References


Finding Cervical Cancer Symptoms in Swedish Clinical Text using a Machine Learning Approach and NegEx

Rebecka Weegar, M.Sc¹, Maria Kvist, MD, PhD¹², Karin Sundström, MD, PhD³, Søren Brunak, PhD⁴, Hercules Dalianis, PhD¹

¹Department of Computer and Systems Sciences, (DSV), Stockholm University, Sweden; ²Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, Stockholm, Sweden ³Department of Laboratory medicine (LABMED), Karolinska Institutet, Stockholm, Sweden ⁴NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

Abstract

Detection of early symptoms in cervical cancer is crucial for early treatment and survival. To find symptoms of cervical cancer in clinical text, Named Entity Recognition is needed. In this paper the Clinical Entity Finder, a machine-learning tool trained on annotated clinical text from a Swedish internal medicine emergency unit, is evaluated on cervical cancer records. The Clinical Entity Finder identifies entities of the types body part, finding and disorder and is extended with negation detection using the rule-based tool NegEx, to distinguish between negated and non-negated entities. To measure the performance of the tools on this new domain, two physicians annotated a set of clinical notes from the health records of cervical cancer patients. The inter-annotator agreement for finding, disorder and body part obtained an average F-score of 0.677 and the Clinical Entity Finder extended with NegEx had an average F-score of 0.667.

Introduction

Cervical cancer

Cervical cancer is one of the most common cancers worldwide¹, frequently affecting young women below age 40. Therefore screening and early detection is essential². While cervical cancer incidence and mortality rates have dropped in countries where screening procedures have evolved into established prevention methods³⁴, the rates in less developed countries are still high.

Cervical cancer, highly mortal in advanced stages, is usually described as having few early symptoms except vaginal discharge, bleeding and pain post coitus. Early detection of cervical cancer is however crucial for treating it successfully. Novel ways to diagnose the disorder at even earlier stages would be highly valuable, as this could prevent treatable pre-cancer from turning into invasive cancer⁴. Early detection is yet sometimes hindered since not all women participate in cervical screening programs.

Today, electronic health records that describe the whole healthcare period of a patient are available in many countries including Sweden. To define and detect these possible early symptoms in a patient's health record, text-mining tools capable of identifying symptoms discriminatory for cervical cancer, are needed.

The overall aim of this study is to evaluate the performance of two previously developed text mining tools on health records of patients with cervical cancer, with a further aim to detect and discover early symptoms in a patient's medical history.

Text mining in the cervical cancer domain

Previous research in the field of text mining and cervical cancer has mostly focused on oncologic documents and pathology reports. One study used text mining of pathology reports, with the aim to transfer unstructured pathology reports to a structured database⁵. A review article⁶ regarding different approaches for clinical text mining within the cancer domain describes only two studies focused on cervical cancer. Both studies concern methods used to retrieve scientific oncology documents relevant to clinical decisions within a particular domain of cervical cancer. None of
them are in the domain of clinical text mining of symptoms relating to cervical cancer. However, there are two articles on the subject of cervical cancer staging. In the first, the authors used 250 cervical patient records for their experiments. The cervical cancer tumors were assigned 15 parameters to classify the stage of the cancer and 0.73 accuracy was obtained using a neural network architecture. In the second article, 221 cervical patient cases were used as input to a staging system using soft computing/neural networks. The cases were classified in stages I-IV and a classification score of 0.79 was obtained on test data. The C4.5 machine-learning algorithm was also applied on the data set, obtaining a score of 0.80.

In another study on tumor staging, the authors trained Naïve Bayes, Bayesian Network, Support Vector Machines, and Random Forest algorithms on manually annotated pathology reports from one hospital and evaluated the portability on another hospital’s pathology reports. They noted a decline in performance with at least 25 percent. However, in a later study they improved their results considerably to only a few percent decline in cross-hospital evaluation, by using feature selections to tune the classifiers and simple rules to identify numeric values.

**Named entity recognition and negation detection**

The Clinical Entity Finder (CEF) is a tool for entity recognition based on the machine-learning algorithm CRF++ and trained on manually annotated clinical entities in Swedish texts from an internal medicine emergency unit. An earlier approach to detect findings and disorders in the cervical cancer domain using Clinical Entity Finder revealed that many of the entities found by the tool were negated. Negations are important in essentially all clinical text, as many diagnoses are made based at least in part on exclusion of certain symptoms or test results. Negation detection have for example been incorporated in adverse drug reaction detection text mining workflows and similarly in the search for diagnostic codes in free text in electronic medical records. The Clinical Entity Finder has no built-in negation detection and a possible tool for negation detection in clinical text is NegEx, a rule based tool. NegEx has previously been adapted to Swedish with a set of Swedish negation triggers.

As it has been shown that different clinical subspecialties use clinical language with different distinctive features, a decrease in the performance of a natural language processing (NLP) system developed in one clinical domain is expected when applied in another clinical domain. The specific aim of this study is to investigate the efficiency of existing tools for text mining of Swedish health records, i.e. for entity recognition and negation detection, respectively, on the subdomain of clinical notes relevant for text mining studies on cervical cancer.

The research questions are:
- How well will a machine learning based tool (Clinical Entity Finder) trained on clinical notes from an internal medicine emergency unit automatically annotate cervical cancer symptoms in physicians’ notes from departments of gynecology and oncology?
- How well can the Swedish NegEx differentiate between negated and non-negated symptoms of cervical cancer?

**Methods and Materials**

**Data**

The Stockholm EPR Corpus is a database containing large quantities of clinical text in Swedish; over 600,000 patient records encompassing over 500 health care units from the Karolinska University Hospital. For this study, all patient records from the departments of obstetrics/gynecology and oncology from the years 2009-2010 with a diagnosis code for malignant neoplasms in the cervix (ICD-10 codes C53.0, C53.1, C53.8 and C53.9), were extracted, resulting in 646 patient records. The extracted data, containing notes from nurses, physicians and other professionals, is called the Cervical Cancer Corpus. Figure 1 shows the age distribution of the patients in the Cervical Cancer Corpus. For this study the physicians’ notes, in total 17,263 notes and 776,719 tokens were used.

---

\[a\] http://code.google.com/p/crfpp/

\[b\] This study has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2014/1882-31/5
Figure 1. Age distribution of patients in the Cervical Cancer Corpus

![Age distribution of patients in the Cervical Cancer Corpus](image)

Annotation

Annotation is the process of marking relevant entities in a text. This is performed manually by domain experts, and the annotation can be used to train a machine learning tool for automatic entity recognition. Manually annotated text is also used for the evaluation of such tools.

In this study, the entity types finding, disorder and body part were annotated. Both findings and disorders can be relevant as symptoms or indications of cervical cancer. The entity type body part is also relevant since it can give the location of symptoms. A finding is an observation made at a certain point in time and it is not necessarily abnormal, it can be a symptom reported by the patient as well as a finding from a medical examination. Disorders are abnormal and can be present even when they are not observable. The annotation performed in this study follows a set of guidelines based on those used when developing the Clinical Entity Finder, with added instructions for annotating negations.

Pre-annotation with the Clinical Entity Finder and NegEx

Pre-annotation is when a tool is used to mark relevant entities in a text before the manual annotation. This pre-annotation can then be used as a timesaving support for the annotators when doing the manual annotation. All physicians’ notes from the Cervical Cancer Corpus were pre-annotated using the Clinical Entity Finder. The Clinical Entity Finder was used to pre-annotate instances of the named entities body part, finding and disorder. To distinguish between negated and non-negated entities, the Clinical Entity Finder was augmented with the tool NegEx, which was used to classify entities of the types finding and disorder in the corpus as negated or non-negated. NegEx uses lists of negation triggers and looks for the presence of those triggers in text surrounding the annotated instances. A trigger can be a word or a sequence of words indicating negation. An example of a sentence with a negation is Patienten har ej smärta (The patient has no pain), where the finding smärta (pain) is negated by the trigger ej. Body parts and words outside of annotations were not investigated for negation.

Manual annotation of the physicians’ notes

Two physicians annotated subsets of the physicians’ notes from the Cervical Cancer Corpus. One physician is experienced in the text mining domain and in annotation (Annotator A), and one physician is an expert in the area of cervical cancer (Annotator B). They received the pre-annotated data, agreed on the interpretation of guidelines, and revised the annotations performed by the Clinical Entity Finder, using the Brat annotation tool. To properly evaluate the tools at least 100 manually annotated notes were considered as necessary to ensure variation and coverage. Annotator A and Annotator B annotated 180 and 100 notes respectively. The annotators removed faulty annotations, changed the span or type of annotations when needed, added missing annotations and marked annotations when negated. Figure 2 shows an example of a text pre-annotated by the Clinical Entity Finder with corrections carried out by Annotator A.
Figure 2. Screenshot of one sentence processed by the Brat annotation tool showing the pre-annotation in green and the manual annotation in blue. (Key phrases are translated to English).

An overview of the annotation pipeline is shown in Figure 3. The 17,263 physicians notes were firstly pre-annotated using the Clinical Entity Finder. In the next step all disorders and findings found by the Clinical Entity Finder were investigated for possible negations using NegEx. Finally, 180 of these pre-annotated notes were randomly selected for manual annotation (subset A). Annotator A corrected all of them, and Annotator B corrected 100 of them (subset B). The annotator agreement was measured on subset B and the performance of the Clinical Entity Finder was measured on subset A.

Figure 3. Annotation pipeline, note that subset B is covered by subset A, both annotators annotated subset B but Annotator A annotated another 80 notes.

Evaluation metrics

The performance of the tools was measured using precision (P), recall (R) and F1-score\(^2\). A high precision indicates that found named entities are correctly classified; a high recall shows that a large portion of named entities in the text are found, and the F1-score is the harmonic mean of the two. To calculate the scores, the number of true positives (TP), true negatives (TN) and false negatives (FN) were counted. A true positive is achieved when a tool has correctly annotated a named entity, a false positive occurs when a tool has incorrectly annotated a named entity, and a false negative when a tool has failed to detect a named entity that is present in the text.

\[
\text{Precision} = \frac{TP}{TP+FP} \\
\text{Recall} = \frac{TP}{TP+FN} \\
F1 = \frac{2*P*R}{P+R}
\]

The inter-annotator agreement between Annotator A and Annotator B was measured using the F1-score. The inter-annotator agreement can be viewed as an indicator of how well two individuals agree on this task and this can be compared to how well a tool can be expected to perform that same task.

Results

Results of the manual annotation

The annotators A and B each annotated 100 physicians’ notes from medical records of cervical cancer patients. Table 1 shows the confusion matrix for the two annotators. The rows show the annotations performed by Annotator A and the columns show the annotations performed by Annotator B. The diagonal shows the numbers of true positives - instances of named entities where the annotators agree on the type and the span of the annotation. To get
a match, the two annotators must pick the same exact span of text. An example of when the annotators have marked different spans is when one annotator marked the span *squamous cell carcinoma in situ* as a disorder but the other only marked the word *squamous cell carcinoma* in the same note. Such scope errors fall into the *No exact match* category, together with the cases of annotations carried out by only one of the annotators, either because of disagreement or by mistake. The level of inter-annotator agreement between Annotator A and Annotator B for exact and also partial match where there is at least one overlapping token (word), is shown in Table 2.

**Table 1.** Confusion matrix for Annotator A and Annotator B obtained from 100 clinical notes

<table>
<thead>
<tr>
<th></th>
<th>Body part</th>
<th>Finding</th>
<th>Disorder</th>
<th>Neg. Finding</th>
<th>Neg. Disorder</th>
<th>No exact match</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Body part</strong></td>
<td>253</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td><strong>Finding</strong></td>
<td>5</td>
<td>256</td>
<td>27</td>
<td>3</td>
<td>0</td>
<td>153</td>
</tr>
<tr>
<td><strong>Disorder</strong></td>
<td>2</td>
<td>3</td>
<td>108</td>
<td>0</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td><strong>Neg. Finding</strong></td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>48</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td><strong>Neg. Disorder</strong></td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>No exact match</strong></td>
<td>36</td>
<td>141</td>
<td>62</td>
<td>16</td>
<td>5</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 2.** F1-scores for Inter-Annnotation Agreement between Annotator A and B. The first row shows the scores when only exact matches are allowed; the second row shows the scores when also including partial matches.

<table>
<thead>
<tr>
<th>Entity type</th>
<th>Body part</th>
<th>Finding</th>
<th>Disorder</th>
<th>Neg. Finding</th>
<th>Neg. Disorder</th>
<th>Negated Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A, B exact match</strong></td>
<td>0.78</td>
<td>0.60</td>
<td>0.62</td>
<td>0.70</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td><strong>A, B partial match</strong></td>
<td>0.79</td>
<td>0.76</td>
<td>0.73</td>
<td>0.81</td>
<td>0.08</td>
<td></td>
</tr>
</tbody>
</table>

**Results of the annotation performed by the Clinical Entity Finder and NegEx.**

All of the physicians’ notes in the Cervical Cancer Corpus were processed by the Clinical Entity Finder and NegEx, and when evaluating the tools, the annotations carried out by Annotator A were used as a Gold Standard, assuming that Annotator A had correctly classified all named entities in the text. Annotator A and the tools were in complete agreement for 54 percent of annotations and Table 3 shows the confusion matrix for Annotator A and the tools, while Table 4 gives the precision, recall and F1-score for each of the entity types. Table 4 also demonstrates the improved scores when allowing partial matches, where there is at least one overlapping token, excluding punctuation.

**Table 3.** Confusion matrix for Annotator A and the Clinical Entity Finder and NegEx combined obtained from 180 clinical notes

<table>
<thead>
<tr>
<th></th>
<th>Body part</th>
<th>Finding</th>
<th>Disorder</th>
<th>Neg. Finding</th>
<th>Neg. Disorder</th>
<th>No exact match</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Body part</strong></td>
<td>285</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>258</td>
</tr>
<tr>
<td><strong>Finding</strong></td>
<td>0</td>
<td>401</td>
<td>6</td>
<td>16</td>
<td>2</td>
<td>375</td>
</tr>
<tr>
<td><strong>Disorder</strong></td>
<td>1</td>
<td>3</td>
<td>181</td>
<td>0</td>
<td>2</td>
<td>62</td>
</tr>
<tr>
<td><strong>Neg. Finding</strong></td>
<td>0</td>
<td>19</td>
<td>1</td>
<td>60</td>
<td>2</td>
<td>47</td>
</tr>
<tr>
<td><strong>Neg. Disorder</strong></td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td><strong>No exact match</strong></td>
<td>29</td>
<td>89</td>
<td>35</td>
<td>5</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 4. Precision, recall and the F1-score for the Clinical Entity Finder and NegEx using annotator A as Gold Standard. The evaluation was performed on 180 clinical notes and the scores for both exact and partial matches are given.

<table>
<thead>
<tr>
<th>Entity type</th>
<th>Body part</th>
<th>Finding</th>
<th>Disorder</th>
<th>Negated Finding</th>
<th>Negated Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>Exact</td>
<td>Partial</td>
<td>Exact</td>
<td>Partial</td>
<td>Exact</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.96</td>
<td>0.78</td>
<td>0.89</td>
<td>0.80</td>
</tr>
<tr>
<td>Recall</td>
<td>0.52</td>
<td>0.55</td>
<td>0.50</td>
<td>0.57</td>
<td>0.73</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.66</td>
<td>0.70</td>
<td>0.61</td>
<td>0.70</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Since NegEx only looks for negations when a finding or disorder has been annotated the performance of NegEx is directly dependent on the performance of the Clinical Entity Finder. The scores given in Table 4 for negated finding and negated disorder are therefore not perfectly suitable for evaluation of how well NegEx performs in the domain of cervical cancer. An additional evaluation has therefore been performed where all findings and disorders annotated both by Annotator A and CEF were selected. By selecting the instances already agreed on, the performance of NegEx could be isolated. There were 500 such instances and of those 57 were marked as negated by both Annotator A and NegEx. Precision, recall and F1-score for the ability of NegEx to correctly determine negations was calculated to be 0.78, 0.75 and 0.76 respectively.

Results of porting the Clinical Entity Finder

Table 5 shows the performance of the Clinical Entity Finder and NegEx in our setting compared to the performance of the Clinical Entity Finder applied to the type of data used to train on; health records from emergency units.

Table 5. The achieved F1-scores of the Clinical Entity Finder combined with NegEx compared to the F1-scores of the Clinical Entity Finder on health records from emergency units.\(^1\)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CEF + NegEx, cervical cancer records</td>
<td>0.66</td>
<td>0.61</td>
<td>0.76</td>
<td>0.47</td>
<td>0.78</td>
<td>0.65</td>
</tr>
<tr>
<td>CEF, emergency unit records</td>
<td>0.85</td>
<td>0.69</td>
<td>0.81</td>
<td>-</td>
<td>-</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Error analysis

A manual error analysis has been performed on 60 of the annotated notes produced by the Clinical Entity Finder and NegEx as compared to the annotations done by Annotator A. The automatic and manual annotation most often classified the annotated entities in the same way, a recurring mistake, however, was made when a compound word for a disorder contained words describing a body part. Such compound words were sometimes wrongly classified as a body part. An example of this is the word *skelettmetastaser* (skeletal metastases), which contains the Swedish word for skeleton. Another source of error is the word *cervixcancer* (cervical cancer), a compound word in Swedish, which in many instances in the records was written as two words, *cervix cancer*. The CEF sometimes interpreted this as two separate entities; the body part cervix and the disorder cancer.

A second step of analysis was performed by looking at instances only annotated by either the automatic or manual annotation. The analysis found 333 false negatives, meaning annotations marked by Annotator A but missed by the tools, and 71 false positives, meaning annotations only performed by the tools. Table 6 shows the different errors sorted by type.
Table 6. The number of false positives and false negatives found in error analysis.

<table>
<thead>
<tr>
<th></th>
<th>Scope Errors</th>
<th>Body part</th>
<th>Finding</th>
<th>Disorder</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positives</td>
<td>47</td>
<td>2</td>
<td>16</td>
<td>6</td>
<td>71</td>
</tr>
<tr>
<td>False Negatives</td>
<td>65</td>
<td>116</td>
<td>132</td>
<td>20</td>
<td>333</td>
</tr>
</tbody>
</table>

After excluding scope errors, about half of the cases of false negatives in all categories were caused by the Clinical Entity Finder failing to detect words and expressions associated with cervical cancer. For example in the category body part, the word cervix was missed several times. The false negatives for disorders directly related to cervical cancer included for example dysplasi (dysplasia) and cancer in situ. Several of the missed disorders were either misspellings or abbreviations, for example the abbreviation cervixca for cervixcancer (cervical cancer in Swedish) and skivepitelcancer a misspelling of skivepitolcancer (squamous cell carcinoma) were both missed by the tools.

The largest source of false positives for disorders and findings were expressions describing procedures or drugs, for example preventivmedel (contraception). For body parts, the false positives originate from expressions where a body part word was part of a non-body part expression. For example, the expression över huvud taget (at all) contains the Swedish word for head (huvud), which causes a faulty annotation.

The individual evaluation of NegEx on 500 instances of findings and disorders resulted in 16 false positives and 19 false negatives. The false positives were almost all due to instances of findings and disorders appearing close to unrelated negation triggers. The false negatives were mainly caused by negation expressions being used that were not included in the lists of negation triggers used by NegEx and there was also a few cases were the manual annotation had failed to discover a negation and borderline cases.

Named entities and negations detected in the physicians’ notes.

Among the 776,719 tokens in physicians’ notes the Clinical Entity Finder identified 58,366 disorders, findings and body parts. Of these, 43,334 were either classified as findings or disorders and a total of 12 percent of findings and disorders were determined to be negated by NegEx. Table 7 shows the most common symptoms and most frequently negated symptoms in the physicians’ notes. The findings and disorders occurring more than 10 times in the text are also sorted on how large a proportion of them that is negated.

Table 7. The most frequent findings, disorders and negations found in the physicians’ notes.
Discussion

In this study, the efficiency of two existing tools for text mining of Swedish clinical text, i.e. for entity recognition and negation detection, was evaluated on a corpus of health records from patients with cervical cancer. The approach involved both manual and automated annotation of clinical entities, including negations. The investigation gave promising results and suggestions for further improvements. The issues studied relates partly to the problem of porting tools between subdomains of clinical text, and partly to the problem of inter-tool dependency when using a pipeline. As a result of this study, lists of clinical entities relevant to cervical cancer were created.

The term “sublanguage” is used to display that language in specialized domains exhibit characteristics that set them apart from general language21. However, clinical language is not homogeneous but consists of several specialized domains that exhibit the characteristics of sublanguages. The subdomain language will influence the creation of NLP tools for clinical text, as a tool relying on term statistics or semantics trained on one clinical note type have been suggested to not work as well on another. Clinical notes from different professions and specialties have been shown to cluster into readily distinguishable groups of lexical and semantic features.7

Here, a machine-learning tool called Clinical Entity Finder trained on clinical notes from a medical emergency unit was applied to notes in a Cervical Cancer Corpus. In both sub corpora, only notes written by physicians were used, as to avoid the further complexity of inter-professional sublanguages. Evaluation by comparing to manual annotation showed that precision was higher than recall for all of the annotated entity types, and manual inspection of tokens classified as findings, disorders, or body parts by the Clinical Entity Finder substantiates the precision of the tool. There were very few apparently faulty classifications; the lower recall was however often related to the tool being unable to classify expressions associated with cervical cancer.

NegEx classifies 12 percent of the findings and disorders in the data as negated; this is line with the findings of the manually annotated English clinical BioScope corpus that contained 14 percent negations. In another annotation study on clinical text from the internal medicine emergency unit, 19 percent of diagnosis expressions were found to be negated. The low F-scores for inter-annotator agreement on negated disorders was partly due to the disagreement between finding/disorder and not only because of disagreement on negation/non-negation. Error analysis did not indicate that the domain of cervical cancer affected the performance of NegEx. Instead the error analysis showed that NegEx gave false positives in some cases where an entity appears close to a negation trigger without actually being affected by it. False negatives where mainly the result of unusual negation triggers used in the physicians’ notes.

Findings and disorders associated with cervical cancer were often negated in the records, this is perhaps due to the fact that the patients in the corpus have known cancers and it is therefore of importance to document the presence or absence of such symptoms. Furthermore, as expected, statements regarding the general well being of the patient such as må brä (feels fine) or oro lig (worried) were frequently present in the records. These kinds of statements were actually seldom negated. Previous studies have found that affirmed and negated findings may be expressed on different levels of hierarchy. Here, this can be exemplified by the findings gynecological problems and signs of infection that were dominantly negated. However, when the patients had such problems, they were described as more fine-grained symptoms, such as bleeding, fever or pain, as opposed to the higher hierarchy general terms.

Manual annotation of clinical text is a time consuming and costly process and there is a need for domain experts or training of non-experts to obtain good quality annotations. Crowdsourcing is not an option due to the sensitive nature of the data. A tool that makes the annotation process easier is therefore valuable. It is also inevitable for human annotators to miss some entities when annotating, and the tools described here may increase the quality of the annotation in such cases.

Future work

Future work includes to extend the training data of the Clinical Entity Finder with the manually annotated cervical cancer health records, as this could be expected to improve the results on the cervical cancer domain by reducing the number of false negatives as indicated by the error analysis. The annotations performed in this study were on a relatively small amount of text. For the retraining of the tool, more annotation is needed, and a Gold Standard needs to be constructed by reviewing the annotations. The error analysis showed that a majority of false positives for
disorders and findings were incorrectly classified procedures or drugs; it could therefore perhaps be useful to include procedures and drugs as entity classes in the annotation to increase the precision of the Clinical Entity Finder. The symptoms found by the tools revealed that the same symptom was written in many different ways. This indicates the need for further usage of stemming, lemmatization and perhaps clustering of the found symptoms, to be able to draw more robust conclusions from our findings.

Conclusion

To our best knowledge, this is the first time cervical cancer narratives have been annotated both by human annotators and a machine learning based tool. By porting the Clinical Entity Finder to this new domain from the domain in which it was trained, the performance was reduced. Precision is still good, but the recall is decreased. This is expected and could be caused by the fact that the new domain contains patterns previously unseen by the tool.

The performance of negation detection was less dependent on the domain change, but since the negation detection is dependent on entities being found by the Clinical Entity Finder, its performance was more difficult to evaluate.

The Clinical Entity Finder was shown to be a useful tool for pre-annotation, since both annotators perceived it as making the annotation process more efficient as compared to annotating text without pre-annotation.

The clinical text found in health records is a promising source of information and the tools evaluated here could be a first step towards finding early symptom patterns in patients with cervical cancer.

Acknowledgements

We thank Maria Skeppstedt for letting us use the Clinical Entity Finder and we also thank Claudia Ehrentraut for her excellent initial study on cervical cancer text mining. This work was supported by the Nordic Information for Action eScience Center (NIASC); a Nordic Center of Excellence financed by NordForsk (Project number 62721). This work was also partly supported by the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by the Swedish Foundation for Strategic Research under grant IIS11-0053.

References


Machine Learning for Treatment Assignment: Improving Individualized Risk Attribution

Jeremy Weiss, PhD1, Finn Kuusisto, MS1, Kendrick Boyd, PhD1, Jie Liu, PhD2, David Page, PhD3

1 University of Wisconsin, Madison, WI 2 University of Washington, Seattle, WA

Abstract
Clinical studies model the average treatment effect (ATE), but apply this population-level effect to future individuals. Due to recent developments of machine learning algorithms with useful statistical guarantees, we argue instead for modeling the individualized treatment effect (ITE), which has better applicability to new patients. We compare ATE-estimation using randomized and observational analysis methods against ITE-estimation using machine learning, and describe how the ITE theoretically generalizes to new population distributions, whereas the ATE may not. On a synthetic data set of statin use and myocardial infarction (MI), we show that a learned ITE model improves true ITE estimation and outperforms the ATE. We additionally argue that ITE models should be learned with a consistent, non-parametric algorithm from unweighted examples and show experiments in favor of our argument using our synthetic data model and a real data set of D-penicillamine use for primary biliary cirrhosis.

Introduction
Estimation of the risk of a disease attributable to an exposure or treatment is an important task in epidemiology and is typically determined using randomized controlled trials (RCTs). The average treatment effect (ATE)–the primary outcome of an RCT–is the average difference between treatment arms in the probability of the outcome, which is then used to recommend future treatments for individual patients. While ATEs are indicative of true treatment effects even in the presence of confounders, they have limited applicability for individual patients because we do not expect the same treatment effect in every person and diversity of effects goes beyond a population’s nonuniform prior risk. In addition, the ATE is population-distribution dependent, so it inherently lacks generalizability to alternative test distributions. Therefore, we consider modeling the individualized treatment effect (ITE), which is the effect of administering the treatment to a particular patient specified by a set of recorded features. Access to the ITE, in addition to the ATE, has many important clinical applications, with just one illustrated below.

Medication Use: Medications almost certainly have different effects in different individuals. For example, hormone replacement therapy treatment effect findings in RCTs and observational studies were of opposite sign for coronary heart disease, and advocacy of their use was rescinded when the RCT findings were released1. Yet, estrogen therapy is still the first line treatment among women experiencing hot flushes. This raises the question of whether ITE modeling can help determine subsets of patients who are still likely to receive benefit. Similarly, many drugs are taken off the market due to excess harm from adverse drug effects. Accurate ITE estimation could bring such drugs safely back to market for select subpopulations.

Currently, non-randomized epidemiological studies adopt classical statistical procedures, such as logistic regression (LR), in seeking to improve patient outcomes. However, machine learning has developed many alternative models for conditional probability distributions (CPDs) with numerous advances achieved. These advances should be leveraged in estimation of the treatment effect—a crucial epidemiological outcome of interest. Our work proposes the use of non-parametric algorithms possessing consistency results in place of logistic regression because of their theoretical ability to accurately recover the CPDs.

In this work, we show the value of ITE over ATE as well as the use of conditional probability models over logistic regression, using both synthetic and real data. We demonstrate our ability to recover the true ITE in synthetic data, and we show the generalizability of the conditional probability model to alternative population distributions of increasing Kullback-Leibler (KL) divergences. We also show that a conditional probability model learned with a consistent, non-parametric algorithm achieves a lower mean squared error (MSE) estimate of the ITE than logistic regression. Furthermore, we show that the conditional probability model produces a better estimate of the ITE than logistic regression on a real RCT data set of D-penicillamine use for primary biliary cirrhosis. Additionally, we show that learning from propensity-score matched examples and stable inverse probability of treatment-weighted examples do not improve over unweighted examples for making ITE prediction when only observational data is available. Thus, by casting treatment effect estimation in a machine learning framework, we introduce ways machine learning can be
used to develop improved, personalized-risk estimates and treatment recommendations.

**Background**
Randomized controlled trials are the gold standard for estimating the average treatment effect. They randomize patients to different treatment arms and measure the rate or probability of an outcome. The treatment arm with the highest success rate determines the preferred treatment, and the conclusion is that future patients who fit the entry criterion of the study should all get the preferred treatment. Randomization is crucial to balance confounders, which are covariates that lead to the outcome and are associated with the treatment. Randomization also balances confounders not measured in the study, so the conclusion is free of confounding bias in expectation.

In general, one cannot observe both “potential outcomes”, i.e., know what will happen to a specific patient under each treatment arm. The treatment that is given elicits the “true” outcome, and the treatment(s) not given elicits the “counterfactual” outcome. The counterfactual outcome is impossible to measure, but with randomization and the assumption that patients are drawn from an underlying population distribution, the expected outcome of patients assigned to a treatment arm is the same as the expected outcome of patients with the same treatment, true or counterfactual. Thus, RCTs provide a recommendation about the treatment effect for every treatment arm in the study for every patient.

**Confounding in Observational Studies:** The RCT, however, is impractical or infeasible for many exposure-outcome pairs. For example, randomization to a harmful treatment, such as smoking, is unethical. In such cases, observational studies are used to derive risk attribution statements. These include studies that use known-confounder-modeling, propensity scoring, inverse probability of treatment-weighting, and doubly-robust estimators. The two main ideas in these methods are to (1) adjust for confounders by modeling them, and (2) manipulate the population distribution so that the treatment is independent of confounders given the outcome. These methods rely upon modeling, but cannot do so effectively if they are missing important contributors to their model: the unobserved confounders. Thus, one key assumption in all of these methods is that there are no unobserved confounders (NUCA), which is difficult to determine in practice. Also, in most of these approaches, a model is assumed for the CPD of the outcome given the exposure and covariate. In these cases, the counterfactual outcomes, which are never observed, are assumed to follow the model CPD.

A second assumption made in clinical studies is the exclusion of intermediate variables–covariates that are on the causal pathway from the treatment to the disease. If included, the treatment effect is underestimated because the effect can be “explained away” by the intermediate variable. The exclusion of intermediate variables decreases the richness of the model, as the intermediate variable may also modify the treatment effect, and analyses that acknowledge and integrate this information exist.

**Individualized Treatment Effect:** One critical drawback of all these methods is that they seek to calculate the average treatment effect, when most applications of risk attribution really desire the ITE. The ITE provides the effect per individual instead of a population-level effect, and information about future individuals can be leveraged in determining optimal treatment choices. Unlike in ATE estimation though, acquiring sufficient counts to estimate the counterfactual ITE outcome is unachievable for any moderate-sized feature vector because the number of possible feature states is exponentially large. Therefore, a modeling approach to estimate the counterfactual outcome becomes necessary. These can be the same CPD models used in pseudo-randomized ATE estimation, e.g. logistic regression, but in the Methods section we will discuss two reasons to adopt other machine learning models: non-uniform treatment recommendations and non-parametric consistency.

The call for adoption of the ITE is not new, and the limitations of applying population-average effects on individuals has been noted. The ATE or relative risk is stated as the primary outcome, usually followed by a secondary analysis of the heterogeneity of treatment effect. As mentioned in Hayward et al., performing subgroup treatment effects is usually more effective in risk-stratified subgroups derived from multivariate analyses than in subgroups defined by individual covariates, and these methods have been adopted for approximating individualized treatment effects. While these methods do improve finer-grained treatment effect estimates, factors beyond the outcome risk may influence the treatment effect and can be utilized when modeling the ITE.

Modeling of the individualized treatment has been implemented in several studies. Qian et al. develop the framework of reward modeling and using model predictions to estimate individualized treatment rules (ITRs). Our work is related but instead makes statements about the utility of the ITE, the generalizability of the ITE, and the preference for using unweighted observational data for ITE estimation, all with simulations to illustrate these advantages. Our simulations
based on synthetic data have access to a ground truth ITE, which we use to assess our ITE estimations.

However, it is possible to assess the benefit of ITE without access to ground truth. Vickers et al.\textsuperscript{13} provides an unbiased method to estimate the advantages of using the ITE recommendation over the ATE recommendation using existing RCT data. They show that by counting outcomes in the subset of patients where ITE- and ATE- treatment recommendations disagree, the expected difference in treatment recommendations is estimated. Our experiments include such analyses to show that the ITE-recommendation can be estimated without access to the counterfactual outcomes. Unfortunately, this method can be severely underpowered in the case that the ITE- and ATE- treatment recommendations are highly similar, and a power calculation analysis to determine recruitment size would be helpful. Alternatively, a new RCT study could be implemented by randomizing to ITE- and ATE- treatment arms.

The methods we adopt do not directly optimize the individualized treatment recommendation. Instead, we model the conditional probability distribution, and then the differences in probability are determined using the estimates for the treatment effect of the true and counterfactual treatments. Zhao et al.\textsuperscript{14} develop a method to directly optimize for the ITR under a surrogate loss function from RCT data. While this method produces individualized recommendations, we believe a model should also provide treatment effect estimates under each treatment arm, because the treatment effect itself is critical information clinically. Also our methods do not require RCT data and scale to multiple treatment arms and factorial designs.

**Methods**

We describe ITE modeling below. Let the ITE for an outcome $y \in \{0, 1\}$ of a patient with features $v$ given treatment $u \in \{0, 1\}$ be the difference in estimates: $p(y = 1 | u = 1, v) - p(y = 1 | u = 0, v)$. The key assumption made in these modeling approaches is that both potential outcomes—the true outcomes $y_{\text{true}}$ and the counterfactual outcomes $y_{\text{cf}}$—come from the CPD model, that is, $p(y_{\text{true}} | u, v) = p(y_{\text{cf}} | u, v) = p(y | u, v)$ for all $u$ and $v$. The interpretation of the ITE is only causal if the no unmeasured confounders assumption (NUCA) is made; otherwise, it is just a statement about the difference in outcome probability given a new patient described by $(u, v)$.

If we have a correctly specified model and NUCA holds, for any new patient, we have their ITE that guides our treatment choice. This statement is notably population-distribution free and thus can generalize to arbitrary population distributions of $(u, v)$. The ATE does not have this characteristic; its calculation is dependent on the distribution of $(u, v)$ so its application should be limited to populations with similar covariate distributions unless the treatment effect is believed to be uniform.

Recalling that the application of the RCT-recommended treatment suggests that every patient should receive that treatment, a logistic-regression-based model similarly provides a uniform decision. Its decision will be in agreement with the sign of the treatment parameter. However, in many cases, and particularly in those where the treatment effect has small magnitude but high variance, the optimal treatment decision is nonuniform. Thus, we adopt machine learning methods which can estimate the CPD while also providing nonuniform treatment choices. In particular, we use AdaBoost because it has consistency results and is a non-parametric learning algorithm\textsuperscript{15,16}. In other words, AdaBoost will recover the correct CPD given enough examples, and will do so regardless of the train $(u, v)$ distribution provided proper support.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATE</td>
<td>applicability, generalizability</td>
<td>clinical trial gold standard</td>
</tr>
<tr>
<td>ITE</td>
<td>hard to estimate in RCT</td>
<td>applicability, generalizability</td>
</tr>
<tr>
<td>RCT</td>
<td>impractical</td>
<td>balanced confounders</td>
</tr>
<tr>
<td>PSM, (s)IPT-W</td>
<td>NUCA, decreased effective sample size</td>
<td>pseudo-randomized</td>
</tr>
<tr>
<td>LR</td>
<td>uniform treatment recommendation</td>
<td>log odds interpretation</td>
</tr>
<tr>
<td>CPD</td>
<td>potential outcomes follow model</td>
<td>mature machine-learning methodology</td>
</tr>
</tbody>
</table>

Table 1: Discussed methodologies with positive and negative characteristics in green and red respectively. ATE average treatment effect, ITE individualized treatment effect, RCT randomized controlled trial, PSM propensity-score matching, (s)IPT-W (stabilized) inverse probability-of-treatment weighting, LR logistic regression, CPD conditional probability distribution, NUCA no unmeasured confounders assumption.
With the adoption of a non-parametric learning algorithm comes the parametric/non-parametric learning trade-off. Parametric models may require smaller sample sizes to learn effectively but are not consistent if misspecified; non-parametric models may require larger sample sizes to achieve good CPD estimates but have arbitrary joint distribution consistency results.

Recall that propensity-score matching (PSM) and (stabilized) inverse probability-of-treatment weighting ((s)IPT-W) are methods to produce pseudo-randomized data for the estimation of the ATE\(^3,4,5\). With ITE as the target statistic, these methods become less desirable. In modeling the CPD, PSM and IPT-W weighting reduce the effective sample size, reducing our numbers for estimation. Thus under the modeling assumption and the goal of modeling ITE, we argue for unweighted CPD estimation. Table 1 compares and summarizes the advantages and disadvantages of study methods related to ATE and ITE estimation.

**Experimental Approach**

In this section, we restate the claims and reasoning in support of the individualized risk framework and provide ways to confirm them experimentally using synthetic data with access to ground truth, or using observational or RCT data.

As already noted, there is a strong argument for estimating the ITE over the ATE because the ITE is applicable for patient-specific recommendations as opposed to ATE-based, population-average recommendations. With correct specification and NUCA, the ITE is also generalizable to arbitrary population distributions, though it is harder to estimate than the ATE. The value of the ITE recommendation can be compared against an alternative–for example, ATE recommendation–using the subset of randomized patients where treatment recommendations differ\(^13\). With these methods, we test the hypothesis of ITE superiority and illustrate the benefits of ITE estimation on synthetic data.

We suggest that, in preference for generalizability of study outcome, the conditional probability distribution \( p(y|u, v) \) should be modeled with non-parametric learning algorithms. That is, our goal should be to learn the correct \( p(y|u, v) \) irrespective of the distribution \( p(u, v) \) because future data distributions \( p'(u, v) \) may be different. Non-parametric learning algorithms achieve independence from \( p(u, v) \) in the limit of increasing data. We use a synthetic dataset to empirically characterize the recovery of the ITE varying the training set data size and compare the performance of parametric and non-parametric learners varying the similarity of train and test set population distributions. We also use a real RCT data set to compare the treatment assignment policies of parametric and non-parametric learners in the presence of a substantial average treatment effect. Additionally, we use synthetic data to compare ITE estimation generalizability for parametric and non-parametric learners when the test set distribution \( p(u, v) \) varies from the training set distribution, though the conditional probability distribution \( p(y|u, v) \) remains the same. We also show experimentally that estimating \( p(y|u, v) \) directly from the original data distribution outperforms analogous estimators from propensity-score-weighting and similarly to stabilized inverse probability-of-treatment weighting.

Finally, we discuss applications of the conditional probability distribution modeling approach. Numerous concerns have been voiced about the appropriateness of observational data as a data source for the effect of treatments because confounding can bias the statistical interpretation. With free reign on the covariate definitions in observational studies, we may have access to highly-correlated or even logically-related covariates, such as "ever smoked" and "current smoker," in our analysis. We opt to include such covariates for richness of representation that can lead to better estimates of \( p(y|u, v) \), but must adapt our interpretation of "intervention" to specified multivariate changes instead of a (univariate) change of treatment state. We discuss several desired conditions when defining the set of "treatment" states and propose methods to provide interpretable recommendations when the space of "treatment" states is large.

**Experiments**

We first define a synthetic model of myocardial infarction (MI) with thirteen variables: age, gender, smoking status, HDL level, LDL level, diabetes, family history of cardiovascular disease (CVD), blood pressure, history of angina, history of stroke, history of depression, statin use, and MI. For simplicity, each variable is binary. The joint probability distribution is defined by the causal Bayesian network in Figure 1 with hand-crafted conditional probability distributions for each variable informed by medical expertise. Observational data are sampled directly from this Bayesian network, while interventions can be simulated by removing incoming edges to the intervention variable and specifying the Bernoulli distribution parameter. Thus, we simulate data from an RCT of statin use by removing the edge from LDL to statin and using a CPD for statins with equal probability of “yes” and “no” to define our RCT data distribution.

The question we seek to answer for our synthetic model is the effect of statin use on heart attack or MI. We test
the per-patient recommendations for or against statin use from logistic regression and boosted trees against the ATE recommendation of always prescribing statins. Testing uses data generated from our synthetic, randomized distribution and evaluation is performed using both the RCT method of Vickers et al. \(^{13}\) and by comparing the predicted ITE to the ground truth ITE calculated exactly from the Bayesian network. Unless otherwise specified, train and test data are generated from the RCT distribution where the LDL to statin edge is removed from the network. When learning our AdaBoost and logistic regression models, we need to ensure the intermediate and confounder assumptions described in the Background section are met. Since we are using synthetic data, there are no unobserved confounders outside our model. However, diabetes is on a causal pathway from statins to MI, so we exclude it from the features available to our models.

With our synthetic model we seek to characterize estimation of the ITE for each method by looking at error modes of each model and producing learning curves for the models as a function of training set size. To test for applicability to an arbitrary test population distribution, we alter the test distributions by changing CPDs for variables with no parents in the causal DAG. Finally, to evaluate ITE estimation from observational data, we use training set data from the observational Bayesian network and compare the estimation from the unweighted training set with estimation using altered data sets via propensity-score matching and stabilized inverse probability-of-treatment weighting.

To validate our claims on real data, we run experiments using trial data for the treatment of primary biliary cirrhosis (PBC) of the liver from the Mayo Clinic \(^{17}\). The trial covered a ten-year period and randomized patients across treatment with a placebo versus treatment with D-penicillamine. The data set includes 16 variables, including demographic

<table>
<thead>
<tr>
<th>Variable</th>
<th>Count (n=288)</th>
<th>%</th>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex female</td>
<td>255</td>
<td>88.5</td>
<td>Age in years</td>
<td>50.6</td>
<td>10.5</td>
</tr>
<tr>
<td>Ascites</td>
<td>24</td>
<td>8.3</td>
<td>Serum bilirubin (mg/dL)</td>
<td>3.3</td>
<td>4.7</td>
</tr>
<tr>
<td>Hepatomegaly</td>
<td>148</td>
<td>51.4</td>
<td>Serum cholesterol (mg/dL)</td>
<td>364.4</td>
<td>224.8</td>
</tr>
<tr>
<td>Spider angiomas</td>
<td>84</td>
<td>29.2</td>
<td>Serum albumin (g/dL)</td>
<td>3.5</td>
<td>0.4</td>
</tr>
<tr>
<td>Edema, untreated or successful</td>
<td>241</td>
<td>83.7</td>
<td>Urine copper (μg/day)</td>
<td>96.1</td>
<td>84.9</td>
</tr>
<tr>
<td>Edema, present despite therapy</td>
<td>27</td>
<td>9.4</td>
<td>Alkaline phosphate (U/L)</td>
<td>2021.1</td>
<td>2213.2</td>
</tr>
<tr>
<td>Histologic stage</td>
<td></td>
<td></td>
<td>Aspartate aminotransferase (U/L)</td>
<td>122.2</td>
<td>58.1</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>5.6</td>
<td>Triglycerides (mg/dL)</td>
<td>124.4</td>
<td>65.4</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>20.8</td>
<td>Platelet count</td>
<td>259.1</td>
<td>96.4</td>
</tr>
<tr>
<td>3</td>
<td>112</td>
<td>38.9</td>
<td>Prothrombin time</td>
<td>10.8</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>34.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
information like age and sex, as well as various lab tests such as serum albumin, serum cholesterol, and triglycerides (see Table 2). The question we seek to answer for this RCT dataset is the effect of D-penicillamine use on three-year survival. For the three-year survival period, the data set is censored to 288 patients, with 146 in the treatment group and 142 in the placebo group. At the end of three years, the treatment group experienced 27 deaths out of 146, whereas the placebo group experienced 32 out of 142. The trial thus demonstrates an average treatment effect of around 4% reduction, indicating a number needed to treat (NNT) of 25, in death rate at three years.

With the PBC trial data, we seek to compare the estimation of the ITE for each method against the ATE recommendation to treat all patients. Furthermore, given the strength of the average treatment effect, we compare the number of times each method suggests that a patient receive no treatment as opposed to the ATE recommendation. We estimate the average ITE versus ATE and number of untreated patients for each method by running 100 replicates.

We use the ‘ada’ package implementation of AdaBoost in R to learn the boosted forest\(^\text{15}\text{,}\text{16}\). Though the consistency guarantees for AdaBoost require the number of iterations to grow linearly with the training set size\(^\text{18}\), we use the
square root of the training set size to reduce the computational burden. Otherwise, we use the default settings from the 'ada' package. Logistic regression models are trained using the 'glm' function of R.

Results

Figure 2 shows the utility of adopting the ITE recommendation over the ATE recommendation on our synthetic model. We want to lower the risk of MI, so a negative difference between ITE and ATE is desirable. From Figure 2a, we see that the adoption of the ITE recommendation, as calculated from the AdaBoost model, lowers the probability of MI by 0.01 on average, provided sufficient training data. That is, the number needed to treat is about 100, so treating 100 patients with the ITE-recommended treatment results in one less MI on average than the ATE-recommended treatment of giving everyone statins. Only AdaBoost is able to provide an improved recommendation because of its ability to accurately estimate the conditional probability distribution. Since there are no interaction terms in our logistic regression model, the recommendation converges to the ATE recommendation of giving everyone statins, resulting in the observed difference of 0 for larger training set sizes in Figure 2.

Figure 2b shows the estimated expected difference in probability of MI between ITE- and ATE- recommended treatments among patients where the recommendations disagree on treatment choice. We see that the ITE recommendation lowers the probability of MI in this subset by about 0.02, or a NNT of 50. The upturn for AdaBoost as the training set size approaches 100,000 is likely due to correctly identifying more patients with small benefits from not taking statins. This dilutes the ITE- and ATE- difference among those patients where the recommendations disagree, but the population-wide probability of MI, which is the primary value of interest, continues to decrease.

Figure 3 shows the utility of adopting the ITE recommendation over the ATE recommendation on the PBC trial data. We want to lower the rate of death over a three-year period, so a negative difference between ITE and ATE is desirable, just as it is with our synthetic model. In Figure 3a we see that neither the AdaBoost model nor the logistic regression model outperform the ATE recommendation to treat all patients. While we would prefer to see the ITE outperforming the ATE, this result is not altogether surprising given the effectiveness of treatment. We do, however, see that the AdaBoost model roughly matches the ATE and outperforms the logistic regression model as we hypothesize and demonstrate repeatedly with our synthetic data. In Figure 3b we show the average number of patients for which each model recommends no treatment. The AdaBoost model rarely recommends no treatment, showing that it has effectively learned the treat-all policy, whereas the logistic regression model frequently recommends no treatment.

Learning curves for logistic regression and AdaBoost are shown in Figure 4a. These curves show mean-squared error of the ITE predictions versus training set size. As we expect, the parametric logistic regression converges to a non-zero error because the model is misspecified (since the ground truth model is not log-linear in the exposure and covariates). The error of AdaBoost, however, continues to decrease towards 0 as training set size increases, showing very accurate estimation of the ITE is possible with sufficient data. AdaBoost's approach toward 0 error is in line with the non-parametric consistency results of Bartlett et al.18.
which is exactly LDL, HDL, and smoking in the logistic regression would improve its performance. AdaBoost has the capability to simulate applying results to different populations by adjusting the prevalence of the five variables in our Bayesian network. Our final experiment is on the generalizability of ITE predictions for AdaBoost compared to logistic regression. We simulate applying results to different populations by adjusting the prevalence of the five variables in our Bayesian network (see Figure 1) with no parents: age, gender, HDL, depression, family history of CVD. We train on our default RCT data, with marginals shown in Figure 1, but test on modified RCT data with different prevalences of the aforementioned variables. Thus, we have changed \( p(v) \), but \( p(y|u, v) \) remains the same, so an accurate prediction of CPD, which is exactly \( p(y|u, v) \), should handle the changing test distribution gracefully. Figure 4b shows that the MSE does not change much.

To further investigate the errors in ITE prediction, we show the predicted ITE versus the ground truth ITE (as calculated from our Bayesian network) in Figure 5. Additionally, we plot the ATE which effectively predicts an identical treatment effect for all patients. The goal is to have predictions as close as possible to the true value, i.e., to have points as close to the \( y = x \) line as possible. In agreement with the results of Figure 4a, AdaBoost makes better ITE predictions than logistic regression and improves noticeably from 10,000 to 100,000 examples. For logistic regression (top), all ITE estimates will be above or below the line \( y = 0 \) because the model assumes that a single coefficient determines the direction of the effect. The four groupings of points extending down (at various angles) from the origin correspond to various settings of the variables LDL, HDL, and smoking. This suggests that including interaction terms among statins, LDL, HDL, and smoking in the logistic regression would improve its performance. AdaBoost has the capability to learn arbitrary interactions and can provide individualized recommendations, though the errors remain greater than zero as shown in the bottom of Figure 5. Indeed, some of the groupings of points lie off the \( y = x \) line also correspond to certain patient subpopulations for which the ITE is consistently misestimated. We expect, due to the consistency of AdaBoost, that these errors will decrease as more training data is available.

The effect of different data-weighting and matching schemes is shown in Figure 6. The recovery of the CPD model, and thus the ITE, requires the fewest examples when leaving the examples unweighted or using stabilized inverse weighting. While propensity score matching produces worse estimates of ITE, there is no benefit for using stabilized inverse weighting over no weighting for this task. One important consideration is that our data set includes some patients without elevated LDL who take statins, motivated by the suggestion that there could be therapeutic benefit of statins even in borderline hypercholesterolemia. However, in a data set with fewer normal-LDL statin users, propensity-score matching and particularly stabilized inverse weighting will impair the CPD model, because it will attach large excess weight to few examples.

Our final experiment is on the generalizability of ITE predictions for AdaBoost compared to logistic regression. We simulate applying results to different populations by adjusting the prevalence of the five variables in our Bayesian network (see Figure 1) with no parents: age, gender, HDL, depression, family history of CVD. We train on our default RCT data, with marginals shown in Figure 1, but test on modified RCT data with different prevalences of the aforementioned variables. Thus, we have changed \( p(v) \), but \( p(y|u, v) \) remains the same, so an accurate prediction of CPD, which is exactly \( p(y|u, v) \), should handle the changing test distribution gracefully. Figure 4b shows that the MSE does not change much.
Figure 6: Learning curves for AdaBoost trained on observational data. Test set ITE mean-squared error as a function of training set size is shown comparing training from unweighted examples (None), propensity-score matched samples with a 1:1 ratio (PSM), and stabilized inverse probability of treatment weighted examples (sIPTW). 95% confidence intervals are calculated over 100 replications.

Discussion and Future Work

Validation on Real Data: Our work highlights the machine learning potential for individualization in clinical scenarios. However, the work we have presented requires empirical justification in several directions. First, while we have provided empirical analysis on one real RCT data set we seek to apply our framework to more clinical data: both RCT and observational. In these settings, we do not have access to the ground truth ITE. Nevertheless, as in Figures 2 and 3, we can adopt the approach in Vickers et al.\textsuperscript{13} to compare outcomes from ITE and ATE recommendations. For rigorous evaluation, we must evaluate on several populations with varying covariate distributions. A characterization of which $p(u, v)$ distributions provide reliable ITE recommendations is critical as well. For example, there may be more uncertainty for patients under-represented in the training set, especially with limited data.

Statistical Theory: In addition, a deeper theoretical understanding of the algorithms and evaluations is necessary. For example, investigation of required training set sizes for non-parametric learning algorithms to outperform parametric algorithms, and logistic regression in particular, would be valuable. A characterization of the number of examples needed to move past the mean squared error of the logistic regression for a given task is important as a factor in determining when we should recommend the non-parametric model ITE instead of the ATE. Additionally, improved statistical understanding of the Vickers et al.\textsuperscript{13} method is critical. Two such statistical properties of particular interest are (1) a measure of the uncertainty in the differences calculated by the Vickers et al\textsuperscript{13} method and (2) power calculations to determine the necessary sample sizes to detect meaningful differences between ATE and ITE.

Epidemiology Methods: Critical factors preventing the automatic use of EHR data for risk attribution are variable definitions, confounders, and intermediate variables. For a pure prediction problem, precisely-defined variables that carry relevant information can improve performance, provided a large enough training set. For risk attribution, however, exclusion of a confounder or inclusion of an intermediate variable can result in biased estimation of both the ITE and ATE. This occurs with ITE prediction from logistic regression or non-parametric learners because the risk is, often arbitrarily, divided between the exposure and the intermediate variable causing inaccurate estimation of the exposure risk. However, we should not simply accept a regime that excludes intermediate variables because inclusion of intermediate variables may enhance our modeling of the conditional probability distribution.

Epidemiological study design often opts for the removal of variables that would actually improve the conditional probability model. We suggest two approaches to explore. First, for interventions, we can define the scope of their influence as a probabilistic change to multiple variables. For example, a diabetes intervention could be represented by the replacement of variable values for “rosaglitazone”, “fasting blood sugar,” and “HbA1c”. Then we can model the effect of intervention by comparing probability of outcomes under intervention or no intervention while richly modeling the conditional probability distribution. Second, a timeline-based analysis where the effect of the intervention...
on apparent intermediate variables, as well as the outcome of interest, can be modeled over time might allow all variables to be leveraged for prediction without hindering risk attribution.

Conclusion

In this work, we illustrated the parallels between the standard clinical study framework designed to determine the ATE and the burgeoning clinical study framework designed to determine the ITE. We highlighted shortcomings of the ATE; first, that the ATE is an average outcome, when in practice we usually care about the ITE for future patients, and second, that the ATE is population-distribution dependent. We then discussed modeling of the ITE. Notably, logistic regression can only recommend one treatment arm if we exclude non-linear and exposure-covariate interaction terms because the coefficient for exposure is either negative or non-negative. Furthermore, unless correctly-specified, logistic regression is not a consistent learning algorithm, so we cannot always recover the true conditional probability distribution, even from large populations. Instead, we adopted another popular framework, boosted trees. We showed that the forest-based ITE outperformed the ATE on a synthetic problem of MI prediction using binary variables, and that the forest-based ITE outperformed logistic regression-based ITE on a real problem of primary biliary cirrhosis treatment with D-penicillamine. Additionally, we showed that the use of propensity-score matching and inverse-probability-of-treatment weighting failed to improve the learning of the conditional probability distribution, suggesting that unweighted samples should be used for learning a model of the ITE. Finally, we showed that the forest-based ITE better generalized to different test set distributions than the logistic regression-based ITE. Modeling of the ITE has large theoretical advantages, though robustness guarantees and validation of its performance should be established before large-scale clinical deployment. A few sources of validation include replication studies and heterogeneity of treatment effect analyses using ITE model strata.

References

Abstract
Family health history (FHx) is one of the most important risk factors for disease. Unfortunately, collection and use of FHx is under-utilized in the clinical setting. Efforts to improve collection of FHx have had minimal impact. A novel approach to collect FHx using social networking capabilities is being explored. We conducted a nationwide survey of 5,258 respondents to 1- assess the interest in using an online social network for FHx, 2- identify if such a tool would have clinical utility, and 3- identify notable trends and potential concerns. We found survey respondents to be very supportive of the proposed approach and interesting trends related to age, education, and race were identified. Results from this survey will be used to guide future research and development of a proposed FHx social network application.

Introduction
Family health history (FHx) is one of the most important risk assessment resources to help clinicians prevent and control disease. By knowing a patient’s heritable risks, a clinician can personalize preventative and risk-mitigating strategies, such as increased screening, prophylactic surgery, risk-reducing therapeutics, and lifestyle changes. FHx is among the strongest risk factors for disease. For instance, patients with three or more first degree relatives with breast cancer have a 4-fold increased risk for breast cancer. Even a single first degree relative with colon cancer yields a 5-fold increased risk. Clinicians empowered with this information can personalize preventative care to reflect the increased risks of the patient. For example, colonoscopies and mammograms could be started at an earlier age and at increased frequency for patients at increased risk for colon or breast cancer, respectively. Furthermore, FHx is becoming a critical component of genomic medicine as the FHx is often used by genetics professionals to interpret genetic test results. Because FHx plays a significant role in risk assessment and genomic interpretation, its importance to controlling and preventing disease is considerable.

Even though FHx is a valuable resource to disease control and prevention, it is largely underutilized in the clinical setting. Studies have shown that clinicians typically discuss FHx with half of new patients and less than a quarter of return patients. Even among patients diagnosed with breast or colon cancer, two diseases known to have strong hereditary components, less than half of clinicians discussed FHx. A good FHx takes about 20-30 minutes to complete; however, clinicians typically only have a few minutes to discuss FHx with their patients. There simply is not enough time during a busy clinic schedule to collect and analyze a 3-generation family pedigree, the gold standard for FHx information. Furthermore, even if a clinician had enough time to record a detailed FHx, a patient’s self-reported FHx is often suboptimal. For example, the accuracy of self-reported FHx ranges widely, from 30%-90%, depending upon the degree of family relatedness and the reported disease. In summary, clinicians lack the time to record what often will be an inaccurately reported FHx from their patients.

Current Approaches to FHx Collection
To overcome these challenges, a number of groups have developed paper-based and web-based tools to help patients organize their FHx outside the clinical visit. The premise of such an approach is two-fold: first, patients have greater access to their FHx information outside a clinic visit, thus allowing them to collect relevant, and hopefully more accurate, FHx without time constraints. Second, and very importantly, when a patient returns to a clinic visit with a completed FHx in hand, the clinician can focus on actually using the FHx information to guide care. The most prominent of these FHx tools is the Surgeon General’s web-based My Family Health Portrait (https://familyhistory.hhs.gov). While small focus groups and surveys have demonstrated that patients and clinicians generally accept the use of such FHx tools, they are poorly adopted outside of research settings. Only about 2% of Americans have ever used a web-based FHx tool. We believe this is largely because the patient alone is responsible to gather, record, and assure the completeness and accuracy of all his or her FHx information with current FHx tools, a substantial burden placed solely on the patient. Moreover, these FHx are usually just basic interfaces to document FHx information; there is little intrinsic incentive to motivate adoption and usage. Also, there is a lack of collaborative support for family members, meaning each family member typically enters redundant family health information in
parallel with other family members- a waste of energy and knowledge. Moreover, even though patient-collected FHx using one of these tools has a fairly good (67%-100%) concordance with a FHx collected by a genetic counselor, this tool does not preclude inaccurate recollection (“I think Aunt Jane had breast cancer, or was it stomach cancer?”). Given the known inaccuracies of patient-reported FHx and the less-than-perfect concordance with in-person FHx interviews, we estimate that only about half of patient-generated FHx information using FHx tools will be correct. Any significant deviation from the true FHx interferes with disease risk assessment and genomic interpretation. Indeed, the issue of FHx reliability and accuracy must be addressed by any FHx collection tool in order to use FHx information effectively to improve patient care. Given the importance of FHx and its inherent challenges, we need more effective and accurate tools to help patients collect their FHx.12

A Novel Approach to FHx Collection

We believe that online social networking is a novel and innovative approach that can improve the collection of FHx. Over the past decade, online social networks have dramatically changed interpersonal communication and spreading information at unprecedented speeds; we plan to leverage benefits of online social networking to improve the collection and use of FHx. Indeed, by directly linking family members together to share health information via an online social network, we believe it will (1) engage more individuals in collecting a FHx as a result of cooperative behavior cascades within a family;1> (2) produce a more complete and accurate FHx because the ‘wisdom of the family’ corrects mistakes and fills in gaps in FHx knowledge;14 and (3) facilitate social support between family members, leading to positive behavior change and healthy decision making.15 While online social networks are available for families (e.g. eFamily.com) and for health (e.g. PatientsLikeMe.com), there are no online social networks that combine families and health. Using social networking to collect FHx is an untapped opportunity to improve the prevention and control of cancer, and may be more effective than current FHx tools in helping individuals collect their FHx. To address this important need, we are developing an online FHx social networking called ItRunsInMyFamily.com (http://ItRunsInMyFamily.com), which will help individuals to link up and share health information with their family members. ItRunsInMyFamily.com will be like Facebook, but designed specifically for sharing health information among relatives. See Box 1 for a description of the features and capabilities of ItRunsInMyFamily.com. Of note, the combination of health information, online social networking, and family relations raises an interesting mix of issues surrounding privacy, laws, and ethics. Nevertheless, efforts are currently underway to further explore these ethical and legal issues needs and determine how to best implement available technologies, techniques, and approaches to protect the security, privacy, and confidentiality of users’ health information.

Box 1. A summary of proposed features and capabilities available within ItRunsInMyFamily.com

End users are critical to the success and utility of any application. It is imperative to demonstrate an interest in ItRunsInMyFamily.com among potential end users. Therefore, we conducted a nationwide survey among the U.S. population to assess relative interest in ItRunsInMyFamily.com, to identify trends, and to highlight potential concerns or issues. Results from this survey will be used to guide future research and development, as well as establish a case for ongoing efforts and support for ItRunsInMyFamily.com.

Methods

We used SurveyMonkey Audience (Palo Alto, CA) to produce a nationwide, demographically balanced representation of the United States population, ages 18 and older. SurveyMonkey Audience, commonly used in academic research, has participants complete a demographic profile which includes sex, race/ethnicity, age, education, income, and regional location of the respondent. This demographic information is then used to randomly recruit a population balanced to match the general U.S. population.1> Additionally, the survey results were weighted by demographics to match the 2013 U.S. Census Bureau Population Survey relative to age, race/ethnicity, sex, and education. To maintain
the integrity of response data, SurveyMonkey Audience limits the number of surveys a member can participate in per week, rewards members with non-cash incentives (charitable donations and sweepstakes entries), and regularly benchmarks the members.

The survey consisted of a total 18 questions. Nine questions were assessing respondents awareness and use of FHx (these results published elsewhere). After a short description of ItRunsInMyFamily.com (see Box 2), seven Likert scale questions assessing the likelihood (not likely, slightly likely, moderately likely, quite likely, and extremely likely) of using and acting upon information in ItRunsInMyFamily.com. Additionally, we asked two Likert scale questions assessing the comfort (not comfortable, slightly comfortable, moderately comfortable, quite comfortable, and extremely comfortable) with planned privacy and security approaches to be used by ItRunsInMyFamily.com (asked after a brief overview of the approaches ItRunsInMyFamily.com will use to protect the privacy and confidentiality of its users [see Box 2]). The complete survey including questions and descriptions can be found at (https://www.surveymonkey.com/s/YHHRW7H). The survey questions and informed consent were approved by the Institutional Review Board (IRB) at the Medical University of South Carolina (MUSC).

Researchers and clinicians at the University of Utah and the Medical University of South Carolina are developing a new social networking website for family health history called ItRunsInMyFamily.com. ItRunsInMyFamily.com will help you gather and organize your FHx by allowing you to enter your own health information and then link up to share health information with your relatives. It is like Facebook, but for your family's health. With ItRunsInMyFamily.com, you will be able to:

- Record and maintain your personal health history all in one place online
- Link up with relatives and see what health concerns they have (or had)
- Comment and discuss health issues that run in your family with your family members
- Receive personalized recommendations based upon your personal and family health risks
- Share your family health history information with your doctor

Box 2. ItRunsInMyFamily.com introductory survey text

Privacy and security of one's health information is often brought up as a potential concern to future users of ItRunsInMyFamily.com. As a result, ItRunsInMyFamily.com developers plan to include a number of safeguards to protect the privacy, security, and confidentiality of its users. For example, ItRunsInMyFamily.com will:

- Prevent unauthorized access to your personal or family health information from individuals, companies, or entities;
- Allow you to control access to your personal health information to specific relatives. For example, you could prevent your Uncle Jack from knowing that you have asthma;
- Allow you to control what is shared about yourself to a family member's doctor;
- Use encrypted transmission and encrypted storage of all information provided; and
- Use HIPAA-compliant servers to store your health information.

Box 3. ItRunsInMyFamily.com privacy & security text

Survey data were analyzed using R 3.1.1 and SAS 9.3 (SAS Institute, Cary, NC). Descriptive statistics were used to summarize the results. Multiple logistic regression was employed in several univariate models to assess differences in covariates with respect to different survey response items of interest. Primary covariates of interest include age, gender, education level, and ethnicity, each weighted to the 2013 US Current Population Survey. For ease of interpretation in regression modeling, the original Likert scale responses were dichotomized to ‘more likely’ (extremely, quite, and moderately) vs ‘less likely’ (slightly and not at all). For direct comparison of scenarios regarding the subset of questions pertaining to “How likely would you be to use ItRunsInMyFamily.com if...?”, a generalized estimating equation (GEE) framework was utilized to control for correlated responses among individuals across the four response variables. We strive for a parsimonious comparison of the scenarios; given the analysis of trends showed relatively equal changes in direction across scenarios within levels of each covariate, the covariate adjusted GEE model was fit without interactions to evaluate how they compared, which provides relatively unbiased inference on population level changes in likelihood of using ItRunsInMyFamily.com across the four scenarios.

Results

A total of 5,258 subjects (of 5,438 invited; a 96.69% participation rate) participated in the survey between October 6 through 17, 2014. The demographics of the respondents are reflective of the U.S. population. Women represent 51.6% of respondents, 21.7% of respondents were between the ages of 18-29, 25.7% between 30-44, 27.1% between 45-59,
and 25.5% over the age of 60. ‘White/Caucasian’ represent 66.4% of respondents, 15.1% were ‘Hispanic/Latino’, 12.3% were ‘Black/African American’. 41.7% of respondents had a high school education or less.

**Would you use ItRunsInMyFamily.com?**

We found that **76.8%** of respondents were likely (7.46% extremely, 18.5% quite, 26.7% moderately, and 24.1% slightly) to use ItRunInMyFamily.com to collect and organize their FHx (baseline scenario). When asked their likelihood of collecting, organizing, and sharing their FHx using ItRunsInMyFamily.com if a doctor encouraged its use, **82.9%** of respondents were likely (13.4% extremely, 26.3% quite, 24.9% moderately, and 18.3% slightly) to use it. Similarly, **82.9%** of respondents were likely (13.1% extremely, 27.4% quite, 25.9% moderately, and 16.5% slightly) to use ItRunInMyFamily.com if a family member invited them to join and share their health information. Interestingly, **87.6%** of respondents would likely (31.9% extremely, 29.1% quite, 17.3% moderately, and 9.4% slightly) use ItRunsInMyFamily.com if they were invited by a family member who was suffering from a health because it might help his/her care or treatment. Notably, nearly two-thirds (61%) of respondents would be ‘extremely’ or ‘quite’ likely to use ItRunsInMyFamily.com in this case. A Chi-square test of weighted frequencies indicates these differences in the likelihood of using ItRunsInMyFamily.com across the various scenarios is statistically significant (p <0.001), indicating the distribution of responses to be different depending on the reasoning given to use the website. See Figure 1 and Table 1 for a summary.

![Figure 1](https://example.com/figure1.png)

**Figure 1.** Responses regarding the likelihood of using ItRunsInMyFamily.com

---

**1319**
Table 1. ItRunsInMyFamily.com survey results

<table>
<thead>
<tr>
<th>Survey questions (All results in percentage)</th>
<th>Extremely likely</th>
<th>Quite likely</th>
<th>Moderately likely</th>
<th>Slightly likely</th>
<th>Not at all likely</th>
<th>No response</th>
</tr>
</thead>
<tbody>
<tr>
<td>If a resource like ItRunsInMyFamily.com were available to you, how likely would you use it to help you collect and organize your family health history?</td>
<td>7.5</td>
<td>18.5</td>
<td>26.7</td>
<td>24.1</td>
<td>23.2</td>
<td>0.0</td>
</tr>
<tr>
<td>...if your doctor encouraged you to use ItRunsInMyFamily.com?</td>
<td>13.4</td>
<td>26.3</td>
<td>24.9</td>
<td>18.3</td>
<td>16.3</td>
<td>0.8</td>
</tr>
<tr>
<td>...if a family member invited you to join ItRunsInMyFamily.com to share your health information with him/her?</td>
<td>13.1</td>
<td>27.4</td>
<td>25.9</td>
<td>16.5</td>
<td>16.0</td>
<td>1.1</td>
</tr>
<tr>
<td>...if a family member who was suffering from a health problem invited you to join ItRunsInMyFamily.com to share your health information because it might help his/her care or treatment?</td>
<td>31.9</td>
<td>29.1</td>
<td>17.3</td>
<td>9.4</td>
<td>11.2</td>
<td>1.2</td>
</tr>
<tr>
<td>If you had family health history information on ItRunsInMyFamily.com, how likely would you be to share this family health history information with your doctor?</td>
<td>40.0</td>
<td>30.8</td>
<td>14.5</td>
<td>7.0</td>
<td>7.7</td>
<td>0.0</td>
</tr>
<tr>
<td>If ItRunsInMyFamily.com provided you a personalized recommendation to speak to your doctor about a health risk, how likely would you be to follow that recommendation and speak to your doctor?</td>
<td>26.0</td>
<td>32.6</td>
<td>21.1</td>
<td>10.6</td>
<td>9.2</td>
<td>0.5</td>
</tr>
<tr>
<td>If a family member encouraged you to speak to your doctor about a health risk, how likely would you be to speak to your doctor?</td>
<td>25.2</td>
<td>34.7</td>
<td>21.8</td>
<td>9.4</td>
<td>7.9</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ItRunsInMyFamily.com used these approaches to protect your privacy, how comfortable would you be…</th>
<th>Extremely comfortable</th>
<th>Quite comfortable</th>
<th>Moderately comfortable</th>
<th>Slightly comfortable</th>
<th>Not at all comfortable</th>
<th>No comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>...recording and storing your health information using ItRunsInMyFamily.com?</td>
<td>12.3</td>
<td>26.5</td>
<td>26.7</td>
<td>16.2</td>
<td>17.2</td>
<td>1.1</td>
</tr>
<tr>
<td>...sharing your health information with family members using ItRunsInMyFamily.com?</td>
<td>12.9</td>
<td>27.5</td>
<td>25.0</td>
<td>16.1</td>
<td>17.0</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Multiple logistic regression models showed some interesting trends relative to age, gender, and ethnicity. See Table 2. We found that compared to the youngest age group (18-29 years old), respondents older than 30 years were at greater odds (30-44 OR: 1.59, p<0.001; 45-59 OR: 1.50, p<0.001; 60+ OR: 1.72, p<0.001) to say they would be more likely to use ItRunsInMyFamily.com. Similarly, increased odds were observed in women compared to men (OR: 1.45, p<0.001), black compared to whites (OR: 2.29, p<0.001), and Hispanics compared to whites (OR: 2.12, p<0.001) with regards to being more likely to use ItRunsInMyFamily.com.
Table 2. Multiple logistic regression results across various outcomes

<table>
<thead>
<tr>
<th>Odds Ratio (95% CI)</th>
<th>If a resource like ItRunsInMyFamily.com were available to you, how likely would you use it to help you collect and organize your family health history?</th>
<th>...if your doctor encouraged you to use ItRunsInMyFamily.com to share your health information with him/her?</th>
<th>...if a family member invited you to join ItRunsInMyFamily.com?</th>
<th>...if a family member who was suffering from a health problem invited you to join ItRunsInMyFamily.com to share your health information because it might help his/her care or treatment?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>30-44</td>
<td>1.59*** (1.35 - 1.87)</td>
<td>1.00 (0.85 - 1.18)</td>
<td>1.35*** (1.15 - 1.60)</td>
<td>1.01 (0.82 - 1.19)</td>
</tr>
<tr>
<td>45-59</td>
<td>1.50*** (1.27 - 1.76)</td>
<td>1.06 (0.90 - 1.25)</td>
<td>1.27** (1.08 - 1.50)</td>
<td>0.99 (0.82 - 1.19)</td>
</tr>
<tr>
<td>60+</td>
<td>1.72*** (1.46 - 2.03)</td>
<td>1.27 ** (1.07 - 1.50)</td>
<td>1.30** (1.10 - 1.54)</td>
<td>0.98 (0.81 - 1.19)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>Female</td>
<td>1.45*** (1.30 - 1.62)</td>
<td>1.26*** (1.12 - 1.41)</td>
<td>1.39*** (1.24 - 1.56)</td>
<td>1.50*** (1.32 - 1.70)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS or Less</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>0.96 (0.85 - 1.08)</td>
<td>1.11 (0.98 - 1.25)</td>
<td>1.15* (1.02 - 1.31)</td>
<td>1.45*** (1.26 - 1.66)</td>
</tr>
<tr>
<td>Graduate</td>
<td>0.95 (0.78 - 1.16)</td>
<td>1.22 (1.00 - 1.50)</td>
<td>1.23* (1.00 - 1.52)</td>
<td>1.90*** (1.49 - 2.43)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White, non-Hispanic</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>2.29*** (1.91 - 2.75)</td>
<td>1.47*** (1.22 - 1.76)</td>
<td>1.33** (1.10 - 1.60)</td>
<td>1.25* (1.02 - 1.54)</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>2.12*** (1.79 - 2.51)</td>
<td>1.36*** (1.15 - 1.61)</td>
<td>1.32** (1.11 - 1.57)</td>
<td>1.11 (0.92 - 1.34)</td>
</tr>
<tr>
<td>Other</td>
<td>0.86 (0.68 - 1.08)</td>
<td>0.82 (0.65 - 1.04)</td>
<td>0.85 (0.67 - 1.08)</td>
<td>0.85 (0.65 - 1.10)</td>
</tr>
</tbody>
</table>

† Univariate multiple logistic regressions across various outcomes adjusting for age group, gender, education, and ethnicity, weighted to the 2013 US Current Population Survey

* Significant at 0.05  
** Significant at 0.01  
*** Significant at 0.001

Interestingly, these differences change with the additional scenarios posed in the subsequent questions. For instance with age, there is a significant difference in likelihood of using ItRunsInMyFamily.com in the baseline scenario that changes to no difference in subsequent scenarios. Race also showed a reduction in difference, though blacks still remained statistically more likely to use ItRunsInMyFamily.com for all scenarios. Conversely, education showed an opposite trend between scenarios with college educated respondents going from slightly less likely in the baseline scenario to significantly more likely in scenarios related to family members, particularly when it would impact a family member’s health. See Table 2. There was little change in difference between scenarios in gender.

These results were further analyzed by fitting subject responses via GEE to estimate the odds of using ItRunsInMyFamily.com with regards to the baseline question and subsequent scenarios. Assuming compound symmetry, within subject correlation between the combination of the four scenarios presented was 0.59. With the baseline question as a reference, a doctor recommendation to use the website increases the likelihood of using the website, with an OR of 1.65 (95% CI: 1.57 - 1.74). A family member recommendation increases the odds of use to a slightly higher degree, with an OR of 1.79 (95% CI: 1.70 - 1.89). However, a family member recommendation because it could potentially help their treatment had the greatest effect on overall willingness to use the website, presenting an OR of 3.30 (95% CI: 3.08 - 3.53). While interactions between the scenario variable with each covariate were independently statistically significant (the result of large sample of complete responses), evaluation of mean plots over the scenarios indicate trends to be relatively parallel in most cases between levels of each of covariate, age, gender,
race, and education. However, the noteworthy trends across scenarios related to age and education level described above were further confirmed. See figure 2.

![Figure 2](image)

**Figure 2.** Generalized estimating equation trend plots for responses by age group and education

**Speaking to your doctor about FHx**

If respondents had FHx information on ItRunsInMyFamily.com, 92.3% of respondents would likely (40.0% extremely, 30.8% quite, 14.5% moderately, and 7.0% slightly) share the information with their doctor. Of note, 71% would be ‘extremely’ or ‘quite’ likely to share their FHx with their doctor. See Table 1. We assessed factors influencing a respondent sharing one’s FHx with his/her doctor. Individuals over the age of 60 (OR: 1.78, p<0.001), women (OR: 1.49, p<0.001), and individuals with a college education (undergraduate OR: 1.48, p<0.001; graduate OR: 1.52, p<0.001) are more likely to sharing FHx information with a doctor. Ethnicity was not found to be a significant factor. See Table 3.

As noted in the introduction, ItRunsInMyFamily.com will also identify and provide personalized recommendations to speak to a doctor about health risks that run in the family. As such, 90.3% of respondents would likely (26.0% extremely, 32.6% quite, 21.1% moderately, and 10.6% slightly) follow recommendations provided by ItRunsInMyFamily.com. Similarly, family members using ItRunsInMyFamily.com can influence one to speak to a doctor, in this case 91.1% of respondents would likely (25.2% extremely, 34.7% quite, 21.8% moderately, and 9.4% slightly) speak to a doctor if encouraged to by a family member. See Table 1. With the exception of Hispanic/Latinos now being significant (OR: 1.35, p<0.01), the same trends among covariates exist as the previous question. See Table 3.
Table 3. Multiple logistic regression results across various outcomes

<table>
<thead>
<tr>
<th></th>
<th>If you had family health history information on ItRunsInMyFamily.com, how likely would you be to share this family health history information with your doctor?</th>
<th>If ItRunsInMyFamily.com provided you a personalized recommendation to speak to your doctor about a health risk, how likely would you be to follow that recommendation and speak to your doctor?</th>
<th>If ItRunsInMyFamily.com used these approaches to protect your privacy, how comfortable would you be recording and storing your health information using ItRunsInMyFamily.com?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age Group</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
</tr>
<tr>
<td>18-29</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>30-44</td>
<td>1.09 (0.88 - 1.33)</td>
<td>1.08 (0.90 - 1.31)</td>
<td>1.08 (0.91 - 1.28)</td>
</tr>
<tr>
<td>45-59</td>
<td>1.17 (0.95 - 1.44)</td>
<td>1.14 (0.95 - 1.37)</td>
<td>0.96 (0.81 - 1.13)</td>
</tr>
<tr>
<td>60+</td>
<td>1.78*** (1.42 - 2.24)</td>
<td>1.55*** (1.27 - 1.88)</td>
<td>1.23* (1.04 - 1.46)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
</tr>
<tr>
<td>Male</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>Female</td>
<td>1.49*** (1.28 - 1.73)</td>
<td>1.48*** (1.30 - 1.69)</td>
<td>1.28*** (1.14 - 1.43)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
</tr>
<tr>
<td>HS or Less</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>1.48*** (1.26 - 1.74)</td>
<td>1.56*** (1.35 - 1.79)</td>
<td>0.97 (0.86 - 1.10)</td>
</tr>
<tr>
<td>Graduate</td>
<td>1.52** (1.15 - 2.01)</td>
<td>1.78*** (1.39 - 2.28)</td>
<td>1.08 (0.88 - 1.32)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
<td><strong>Odds Ratio (95% CI)</strong></td>
</tr>
<tr>
<td>White, non-Hispanic</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
<td>1.00 (ref)</td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>1.37* (1.07 - 1.75)</td>
<td>1.60*** (1.28 - 1.99)</td>
<td>1.29** (1.07 - 1.55)</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>1.04 (0.84 - 1.29)</td>
<td>1.35** (1.11 - 1.63)</td>
<td>1.13 (0.95 - 1.34)</td>
</tr>
<tr>
<td>Other</td>
<td>0.93 (0.68 - 1.08)</td>
<td>0.95 (0.72 - 1.24)</td>
<td>0.75 *(0.59 - 0.94)</td>
</tr>
</tbody>
</table>

† Univariate multiple logistic regressions across various outcomes adjusting for age group, gender, education, and ethnicity, weighted to the 2013 US Current Population Survey
* Significant at 0.05
** Significant at 0.01
*** Significant at 0.001

Discussion

The scope of this study was to demonstrate interest and identify trends among potential users for ItRunsInMyFamily.com. We were encouraged by the relative interest and likelihood use of ItRunsInMyFamily.com by potential end users, particularly under specific scenarios. Our nationwide survey demonstrated that a majority of the U.S. population would likely use ItRunsInMyFamily.com to collect and record their FHx if given the opportunity. This is especially true if a doctor encourages its use, if a family member invites them to join, and most strongly if a family member with a health condition invites one to join because it could help that person’s care and/or treatment. This insight will guide development of ItRunsInMyFamily.com to emphasize clinician and family mediated invitations, to encourage sharing of health information, and to provide care recommendations. Importantly from a clinical perspective, we identified that users of ItRunsInMyFamily.com will be likely share FHx information with his/her clinician. Furthermore, many would follow recommendations provided by ItRunsInMyFamily.com or a family member to speak to a clinician about a particular health issue. These findings are fundamental to the clinical utility of ItRunsInMyFamily.com. Finally, we were encouraged by the relative comfort expressed by respondents for the proposed privacy and security approaches. While further research is warranted, particularly to better understand of individuals concerns, these results provide an early indication that privacy and security concerns may not significantly hinder the adoption of ItRunsInMyFamily.com among patients. Nevertheless, these results suggest that ItRunsInMyFamily.com is an acceptable way for patients to collect and use FHx to improve their care.

Notable trends

Some of the trends in the analysis were particularly interesting. Notably, we were surprised that younger individuals were initially less inclined to use ItRunsInMyFamily.com. Although this difference dissipated under various scenarios,
we had initially expected to see higher likelihood of usage among younger individuals because it is an online social networking tool. While this can be explored further, perhaps age and experience of older individuals with health care reinforces to them the need for resource like ItRunsInMyFamily.com. This assumption is reinforced by the observation that the over 60 age group would be statistically more likely to share their FHx information with their healthcare provider over all other age groups. It is important to note that the survey was conducted over the web, implying that respondents across all ages are biased to be more internet savvy. Therefore, it might not represent a true representation of the actual population, particularly among older individuals in this case. Nevertheless, the trend remains interesting and requires further exploration.

We were intrigued by the trends associated with race, particularly among African Americans respondents. Prior studies have shown that African Americans are less likely to participate in recording their FHx in writing. However, our results show an opposite trend with African Americans being more likely to use ItRunsInMyFamily.com and being more comfortable with ItRunsInMyFamily.com privacy and security approaches compared to the other ethnic groups. This may be due to the fact that, in general, African Americans have a very tight-knit extended family and extensive kinship networks on which they rely for informal social support. Other work has shown that the opinions of family members are valued by African Americans when making health care decisions and hence they may be likely to be comfortable with sharing information within the family. Also, previous research has shown that African Americans are more likely to 1) get health care information online, 2) use email to communicate with family members, and 3) use social networking systems on their phones as a way to be connected with family and friends. Therefore, despite the documented health disparities among African Americans, perhaps leveraging social networking technology to collect FHx is an ideal way to overcome health disparities related to FHx.

Future direction

Several efforts are currently underway to further explore various aspects of ItRunsInMyFamily.com prior to development of the social network. Currently, we are conducting a research project to explore ethical and legal issues related to sharing family health information with relatives using an online social network. While these issues have not escaped our consideration, the legal experts and ethicists working on this project do not believe there are any scenarios that would preclude ongoing development and research on the overall project. Rather, they affirm that thoughtfully developed policies and user privacy controls can mitigate many ethical and legal concerns. Indeed, it is within the scope of the project to define a legal and ethics framework which will guide ItRunsInMyFamily.com development and use. In a separate effort, we are identifying relevant health IT data models, standards and terminologies to be implemented within ItRunsInMyFamily.com to promote health information exchange with electronic health records, personal health records, and other health IT applications. By aligning with Meaningful Use requirements and participating in the nationwide health IT infrastructure, it will allow ItRunsInMyFamily.com users to import and export health information with various health IT applications. Continual user engagement is critical to the success of ItRunsInMyFamily.com, therefore we plan to incorporate proven strategies that promote user remuneration, influence, belonging, and significance to establish and maintain user engagement. Finally, once development of the site is complete, we will need to validate that the FHx information collected using an online social network is more complete and accurate than current FHx approaches, as hypothesized. To perform this validation, we plan to compare FHx collected through various approaches (e.g. genetic counselor, My Family Health Portrait, and ItRunsInMyFamily.com) to the health pedigrees available in the Utah Population Database (UPDB). The UPDB is a research database that provides a true gold standard of FHx information because it combines genealogical records, medical records, and other vital records of Utah citizens.

Conclusion

Results from this nationwide survey demonstrate that potential users are supportive of using an online social network to collect FHx. Results from this survey provide justification for ongoing efforts and will inform future research and development.

Acknowledgements

BMW and JS are the founders of ItRunsInMyFamily.com.
References

22. Social Networking Fact Sheet | Pew Research Center’s Internet & American Life Project.
A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text

Yonghui Wu, Ph.D.1, Jun Xu, Ph.D.1, Min Jiang, M.S.1, Yaoyun Zhang, Ph.D.1, Hua Xu, Ph.D.1

1School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Clinical Named Entity Recognition (NER) is a critical task for extracting important patient information from clinical text to support clinical and translational research. This study explored the neural word embeddings derived from a large unlabeled clinical corpus for clinical NER. We systematically compared two neural word embedding algorithms and three different strategies for deriving distributed word representations. Two neural word embeddings were derived from the unlabeled Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II corpus (403,871 notes). The results from both 2010 i2b2 and 2014 Semantic Evaluation (SemEval) data showed that the binarized word embedding features outperformed other strategies for deriving distributed word representations. The binarized embedding features improved the F1-score of the Conditional Random Fields based clinical NER system by 2.3% on i2b2 data and 2.4% on SemEval data. The combined feature from the binarized embeddings and the Brown clusters improved the F1-score of the clinical NER system by 2.9% on i2b2 data and 2.7% on SemEval data. Our study also showed that the distributed word embedding features derived from a large unlabeled corpus can be better than the widely used Brown clusters. Further analysis found that the neural word embeddings captured a wide range of semantic relations, which could be discretized into distributed word representations to benefit the clinical NER system. The low-cost distributed feature representation can be adapted to any other clinical natural language processing research.

Introduction

Clinical Named Entity Recognition (NER) is to identify boundaries and determine semantic classes (e.g., problems, treatments and lab tests) of clinical concept mentions in clinical text. Over the last few years, much attention has been focused on the clinical NER as it’s the critical step of unlocking important healthcare information from narrative clinical text. Much of the important patient information is locked in the narrative clinical text, which is not directly accessible for clinical applications that rely on structured data. Clinical NER systems identify clinical entities from narrative patient reports to support clinical and translational research. Various clinical NER modules have been developed in general clinical Natural Language Processing (NLP) systems (e.g., MedLEE1, MetaMap2, KnowledgeMap3 and cTAKES4). Most of the existing clinical NLP packages are rule-based systems that rely on comprehensive medical vocabularies. Recently, the clinical NLP community organized a series of open challenges with focus on identifying clinical entities from narrative clinical notes, including the 2009 i2b2 (the Center of Informatics for Integrating Biology and the Bedside) challenge5 on medication information extraction, the 2010 i2b2 challenge6 on recognizing medical problems, treatments, and tests entities, 2013 Share/CLEF challenge7 on disorder mention recognition and normalization, and the 2014 SemEval challenge8 on disorder mention recognition and normalization. Researchers developed rule-based systems, machine learning based systems as well as hybrid systems during the challenges. Currently, most of the state-of-the-art clinical NER systems are primarily based on the machine learning models.9,11

Supervised machine learning methods approach the NER as a sequence labeling problem, which aims to find the best label sequence (e.g., BIO format labels) for a given input sequence (individual words from clinical text). Researchers have applied various supervised machine learning algorithms, including Conditional Random Fields (CRFs)12, Maximum Entropy (ME), and Structural Support Vector Machines (SSVMs)13, to recognize clinical entities. Among the supervised machine learning algorithms, the CRFs is the most popular one for NER tasks as it’s intrinsically designed for sequence labeling problem by modeling the relationships between neighbor tokens. A number of top-ranked NER systems are primarily based on the CRFs. The supervised machine learning algorithms work well as researchers manually extract useful features and feature combinations through feature engineering. Orthographic information (e.g., capitalization of letters, prefix and suffix), syntactic information (e.g. POS tags), n-gram information, semantic information (e.g., UMLS concept unique identifier) and disclosure information (sections in the clinical notes) are often used as features in the typical NER systems. The combination of features, such as the...
word combined with POS tags, also prove to be useful. Subsequently, it was identified that the performance of the supervised machine learning algorithm could be further improved by “unsupervised features”, which are typically derived from unlabeled corpora using unsupervised machine learning methods such as Brown clustering. Conventionally, Brown clusters are converted into symbolic IDs to form the unsupervised feature representations. Brown clusters have been used successfully in a number of top-ranked clinical NER systems, such as the system from Bruijin in i2b2 2010 challenge and the system from Zhang et al. from the SemEval 2014 Challenge. The study from Tang et al. also found that unsupervised features could improve the identification of clinical entities that not covered by the training corpus. However, this one-hot word representation has limitations in that it only captures a single aspect relation of a word using sparse binary vectors. Researchers have explored the distributional semantics models to derive distributional word representations. Jonnalagadda et al. explored the random indexing model and found that the distributional word representations could enhance the performance of clinical concept extraction. Henriksson et al. further combined the distributional word representations with a large unlabeled in-domain corpus to generate additional features for de-identification of health records.

Recently, there has been an increasing interest in training word embeddings from large unlabeled corpora using neural networks. Word embeddings are typically represented as a dense real-valued low dimensional matrix \( M \) of size \( V \times D \), where \( V \) is the vocabulary size and \( D \) is the predefined embedding dimension. Each row of the matrix is associated with a word in the vocabulary and each column of the matrix represents a latent feature. The distributed word representations can be derived from the word embeddings. Different from the one-hot word representations such as the clustering feature from Brown clusters, the word embeddings have real-valued numbers to describe multi-aspect relations between words. Usually, the word embedding matrix is first initiated with random values and then tuned using neural networks induced by the neural language model. Bengio and Mikolov proposed different neural networks to train the word embeddings, where the probability of a word given by the previous word was estimated using the cross-entropy criterion. In 2010, Collobert et al. proposed a new neural language model to train word embeddings using ranking loss criteria with negative sampling. The experimental results showed that the ranking based word embeddings derived from the entire English Wikipedia corpus greatly helped the NER task in general English domain.

Previous studies have shown that the neural word embeddings could represent abundant semantic meanings and capture multi-aspect relations into a real-valued matrix. However, there is no conclusion on how to use the real-valued word embeddings in machine learning based clinical NER systems. In the biomedical literature domain, Tang et al. conducted a study to evaluate the different types of unsupervised word representations in biomedical NER task. They used the popular word2vec package to generate the word embeddings and showed that the word embedding features improved the F1-score of a baseline NER system by 0.49% (from 70.0% to 70.49%). The Brown cluster features improved the F1-score by 1.2% (from 70% to 71.2%), which was superior to the word embedding features. Tang’s study directly used the real values from the embedding matrix as features in a CRFs model without any discretization and the corpus size was relatively small (20,000 sentences from BioCreAtIvE II GM corpus and 22,402 sentences from JNLPBA corpus). Recently, research from Wang and Manning showed that conventional supervised machine learning models, such as the CRFs, have a preference for high dimension discrete feature space instead of low dimension real-valued feature space. Later in 2014, Guo et al. proposed two new strategies for deriving distributed feature representations from neural word embeddings trained from the entire English Wikipedia corpus. The experimental results showed that the proposed binarized embedding features (BinEmb – there are three possible values: “positive”, “negative” and “neutral” in BinEmb feature. However, we keep using this name to make it consistent with the previous research) and the distributed prototype features (ProtoEmb) were comparable to the Brown clusters.

However, until now there is no report of using neural word embeddings in the clinical domain. Compared with general English text, the clinical texts are much noisy with frequently occurred ungrammatical sentences, misspellings and abbreviations. It is not clear how the supervised machine learning based clinical NER systems could benefit from the neural word embeddings derived from the noisy clinical corpora. It’s also not clear which of the neural word embedding algorithms would be better for clinical NER tasks and how to utilize the word embeddings as features in machine learning based NER systems. In this study, we propose to 1) explore the power of a large unlabeled clinical corpus (403,871 notes) using deep neural networks (DNN); 2) compare the entropy-based neural word embedding algorithm and the ranking-based word embedding algorithm on the clinical NER task; 3) compare three different strategies for deriving distributed word representations from word embeddings in clinical NER tasks. To the best of our knowledge, this is the first study of training neural word embeddings from a large unlabeled clinical corpus and comparing different neural word embedding algorithms and strategies for deriving
distributed word representations. The most related study is that of Guo et al. in the general English domain, where they used only the word embedding algorithm implemented in word2vec. Another related study is by Tang et al. for NER in biomedical literature, where the study directly used the real values as feature weights without discretization, and the corpus size was relatively small.

Methods

Data sets

This study used the annotated corpora from the 2010 i2b2 challenge and the 2014 SemEval challenge. The i2b2 2010 corpus is annotated with three types of clinical entities including Problem, Test, and Treatment. All entities are composed of consecutive words. The SemEval corpus has only one type of entity – the disorder mention. However, the SemEval corpus contains disjoint entities – entities that are composed of more than one piece of text region. The following sentence illustrates an example: “The aortic root and ascending aorta are moderately dilated”. There are two disjoint entities: “aortic root ... dilated” and “ascending aorta ... dilated”. Table 1 shows the detailed information for the two labeled clinical corpora. In order to train the neural word embeddings and the Brown clusters, we utilized a 2.2 gigabytes of unlabeled clinical notes from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II corpus. The MIMIC II corpus is composed of 403,871 notes from four different note types. Table 1 shows the detailed information about the MIMIC II corpus. All the clinical notes were preprocessed using the same pipeline to separate sentences and tokens.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Notes</th>
<th>Entities</th>
<th>Entity types</th>
<th>Note types</th>
</tr>
</thead>
<tbody>
<tr>
<td>i2b2 2010</td>
<td>Training: 349</td>
<td>27,837</td>
<td>Problem, Treatment Test</td>
<td>Discharge, Progress</td>
</tr>
<tr>
<td></td>
<td>Test: 477</td>
<td>45,009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SemEval 2014</td>
<td>Training: 298</td>
<td>11,156</td>
<td>Disorder</td>
<td>Discharge, radiology ECG, ECHO</td>
</tr>
<tr>
<td></td>
<td>Test: 133</td>
<td>7,971</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIMIC II</td>
<td>N/A</td>
<td>403,871</td>
<td>N/A</td>
<td>Discharge, radiology ECG, ECHO</td>
</tr>
</tbody>
</table>

ECG: electrocardiogram, ECHO: echocardiogram

The Machine learning-based NER framework

To apply machine learning algorithms to the NER task, the annotated corpora are typically converted into “BIO” format, where “B” denotes “the beginning of a concept”, “I” denotes “inside of a concept”, and “O” denotes “outside of a concept”. E.g., a concept for the semantic category of medical problem can be represented using “B-problem” and “I-problem”. In this study, we developed a baseline clinical NER system based on the CRFs model. The baseline system covers the most common NER features including bag-of-word, Orthographic information (word patterns, prefixes and suffixes), syntactic information (POS - part of speech tags), n-gram of word and POS tags (unigrams, bigrams, and trigrams), disclosure information (sections and note types) and combination of words and POSa tags. We used the implementation of CRFs in the CRFsuite package (http://www.chokkan.org/software/crfsuite/). The model parameters were optimized using 5-fold cross validation on the training data, and the best parameters were used to predict the test data.

Integrate word embeddings with the NER framework

Neural word embedding algorithms

We explored two popular word embedding algorithms, including the word2vec from Mikolov and the ranking-based neural word embedding algorithm from Collobert. For word2vec, we used the implementation from “https://code.google.com/p/word2vec/” with the default settings (we used the CBOW model, which is faster and a little bit better than the skip-gram). As there is no out of shelf package for the ranking-based neural word embedding algorithm, we implemented the deep neural network according to the paper from Collobert using Java. We used the suggested parameters to train the neural network with a hidden layer size of 300, a fixed learning rate of 0.01, and an embedding dimension of 50. The standard stochastic gradient descent algorithm was used to optimize the neural network according to the ranking loss. The final word embeddings were represented as dense real-valued matrix. Each row in the embedding matrix associated with a word. Figure 1 shows examples of the word embeddings.
Figure 1. Examples of neural word embeddings

Neural word embedding as features

This study compared three different strategies of deriving distributed word representations from neural word embeddings. For each of the strategies, we derived corresponding distributed representations from the MIMIC II corpus and tested the effect of the derived features using the 2010 i2b2 data and the 2014 SemEval data.

1) Raw embedding feature (RawEmb)

The raw embedding feature is a straightforward way of using neural word embeddings. In this method, the real values from the embedding matrix were directly used as feature weights without any post processing. This method will generate the same number (equals to the dimension of the word embeddings) of feature representations for each word. Tang et al. used this strategy in their research\textsuperscript{14}. Figure 2 shows examples of the raw embedding features.

Figure 2. Examples of raw embedding features

2) Binarized embedding feature (BinEmb)

The binarized embedding feature was proposed by Guo et al.\textsuperscript{26} in 2014 for general English domain. The intuition of the binarized embedding feature is to discretize the real-valued matrix and omit the insignificant dimensions. Given a real-valued neural word embedding matrix $M_{F,D}$, the binarized embedding features can be derived by converting the real-valued embedding matrix to another discrete-valued matrix $M'_{F,D}$ with the discrete symbolic values in $[+,-,0]$. For the $j^{th}$ dimension (column) of the embedding matrix, we first calculate the positive mean $MEAN(j)^+$ and negative mean $MEAN(j)^-$ according to the following equations:

$$MEAN(j)^+ = \frac{1}{N_j^+} \sum_{i=0}^{N_j} M_{i,j}, \ M_{i,j} > 0$$

$$MEAN(j)^- = \frac{1}{N_j^-} \sum_{i=0}^{N_j} M_{i,j}, \ M_{i,j} < 0$$

where $N_j^+$ is the total number of rows with $j^{th}$ column $M_{i,j} >0$, and $N_j^-$ is the total number of rows with $j^{th}$ column $M_{i,j} <0$. Then the discrete-valued matrix $M'$ can be derived by the following projection:

$$M'_{i,j} = \begin{cases} +, & \text{if } M_{i,j} > MEAN(j)^+ \\ - , & \text{if } M_{i,j} < MEAN(j)^- \\ 0, & \text{otherwise} \end{cases}$$

Using the discrete-valued matrix $M'$, we can add the symbolic features according to the row vector associated with the word. Figure 3 shows examples of the binarized embedding features.

Figure 3. Examples of binarized embedding features

3) Distributed prototype feature (ProtoEmb)

The distributed prototype feature was also proposed by Guo et al.\textsuperscript{26} for the general English domain. Instead of using the embedding dimensions, the prototype feature method selects prototypical words as representative features for each label. The prototypical words will be assigned as features to the samples according to the distributed similarity
in the embedding matrix. Typically, the prototypical feature words are selected using the normalized pointwise mutual information (PMI) between the word and its labels (equation 4 and 5). For each word 'w' in the training/test, we calculate the cosine similarity between 'w' and all the selected prototypical words using the associated embedding vectors. If the cosine similarity is above the predefined threshold, the prototypical word will be assigned as a feature. Following the study by Guo\textsuperscript{26}, we tested different numbers of prototypical words and cutoff thresholds to select the best parameters. Finally, the top 40 prototypical words and a cutoff threshold of 0.5 were used to assign the prototypical words for the distributed prototype feature. Figure 4 shows the prototypical feature words selected for each label from i2b2 corpus and Figure 5 shows examples of assigned prototypical words.

\[ nPMI(label, word) = \frac{PMI(label, word)}{-\ln p(label) + p(word)} \]  

(4)

\[ PMI(label, word) = \ln \frac{p(label, word)}{p(label)p(word)} \]  

(5)

**B-Problem:** afibrile acute hypertension vomiting chills nausea nontender chronic mild some moderate ...

**I-Problem:** disease pain breath failure fibrillation mellitus stenosis effusion infarction distress ...

**B-Treatment:** coumadin lisinopril metoprolol protonix aspirin colace heparin tylenol percocet ...

**I-Treatment:** therapy catheter sulfate drip graft scale bypass saline fluids replacement tube support ...

**B-test:** auscultation glucose hgb abs bun wbc rbc mcph plt mcv ast creat cl pt tt creatinine ...

**I-test:** scan count pressure x-ray ct cultures culture saturation rate biopsy exam bilirubin study fraction ...

**O:** : was to with for is she he on and no mg by day as had discharge he has were date history patient ...

Figure 4. Examples of prototype words selected using normalized PMI

**focal : pleural obstructive dependent acute chronic mild moderate metastatic ...**

**warfarin : coumadin lisinopril metoprolol protonix aspirin colace heparin tylenol percocet ...**

**pt : patient she he and has which but also there that but chf wbc pain inr ...**

Figure 5. Example of prototype features assigned to word

**Experiments and Evaluation**

We ran two neural network embedding algorithms and Brown clustering algorithm on the unlabeled MIMIC II corpus to derive word embeddings and Brown clusters. For Brown clustering, we used the implementation from “https://github.com/percyliang/Brown-cluster/” and set number of clusters to 1,000. (We tested the different number of clusters from 50 to 2000 during the 2014 SemEval challenge. The cluster number 1,000 achieved the best performance) For each neural word embedding algorithm, we compared three different strategies for deriving word embedding features. Finally, we combined the best word embedding feature with the Brown clusters to examine how the clinical NER system could benefit from the large unlabeled clinical corpus. The official evaluation scripts provided by the i2b2 organizers and SemEval organizers were used to calculate the strict micro-averaged precision, recall, and F1-score. We report the performances of combining different types of unsupervised word representations on the 2010 i2b2 data and the 2014 SemEval data.

**Results**

Table 2 and Table 3 show the performances of the CRFs based NER system on the 2010 i2b2 data and the 2014 SemEval data respectively, when using different word representation features. The baseline system achieved F1-scores of 0.799 for the i2b2 data and 0.754 for the SemEval data. The baseline performance on SemEval data is lower than the performance on i2b2 data. The Brown cluster features improved the baseline system by 1.7% for the i2b2 data and 1.3% for the SemEval data. The binarized embedding features outperformed other embedding features and the Brown clusters, by improving the F1-score by 2.3% for the i2b2 data and 2.4% for the SemEval data. The combined feature from the binarized embedding features and the Brown clusters improved the F1-scores by 2.9% for i2b2 data and 2.7% for SemEval data, respectively.
Table 2. Results on the 2010 i2b2 data set.

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline features</td>
<td>0.848</td>
<td>0.755</td>
<td>0.799</td>
</tr>
<tr>
<td>+RawEmb (ranking)</td>
<td>0.848</td>
<td>0.768</td>
<td>0.806</td>
</tr>
<tr>
<td>+BinEmb (ranking)</td>
<td>0.849</td>
<td>0.797</td>
<td>0.822</td>
</tr>
<tr>
<td>+ProtoEmb (ranking)</td>
<td>0.849</td>
<td>0.786</td>
<td>0.816</td>
</tr>
<tr>
<td>+RawEmb (word2vec)</td>
<td>0.847</td>
<td>0.766</td>
<td>0.804</td>
</tr>
<tr>
<td>+BinEmb (word2vec)</td>
<td>0.846</td>
<td>0.790</td>
<td>0.817</td>
</tr>
<tr>
<td>+ProtoEmb (word2vec)</td>
<td>0.852</td>
<td>0.782</td>
<td>0.815</td>
</tr>
<tr>
<td>+BrownCluster</td>
<td>0.847</td>
<td>0.788</td>
<td>0.816</td>
</tr>
<tr>
<td>+BrownCluster + BinEmb (ranking)</td>
<td>0.851</td>
<td>0.806</td>
<td>0.828</td>
</tr>
</tbody>
</table>

Table 3. Results on the 2014 SemEval data set.

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline features</td>
<td>0.782</td>
<td>0.727</td>
<td>0.754</td>
</tr>
<tr>
<td>+RawEmb (ranking)</td>
<td>0.775</td>
<td>0.758</td>
<td>0.767</td>
</tr>
<tr>
<td>+BinEmb (ranking)</td>
<td>0.781</td>
<td>0.774</td>
<td>0.778</td>
</tr>
<tr>
<td>+ProtoEmb (ranking)</td>
<td>0.784</td>
<td>0.748</td>
<td>0.766</td>
</tr>
<tr>
<td>+RawEmb (word2vec)</td>
<td>0.778</td>
<td>0.750</td>
<td>0.764</td>
</tr>
<tr>
<td>+BinEmb (word2vec)</td>
<td>0.779</td>
<td>0.764</td>
<td>0.771</td>
</tr>
<tr>
<td>+ProtoEmb (word2vec)</td>
<td>0.789</td>
<td>0.752</td>
<td>0.770</td>
</tr>
<tr>
<td>+BrownCluster</td>
<td>0.778</td>
<td>0.756</td>
<td>0.767</td>
</tr>
<tr>
<td>+BrownCluster + BinEmb (ranking)</td>
<td>0.783</td>
<td>0.780</td>
<td>0.781</td>
</tr>
</tbody>
</table>

All the unsupervised word representation features (including Brown clusters and word embeddings) improved the performances of the clinical NER systems. Further analysis found that the performance improvements are mainly from the recall. For example, the combined features from word embeddings and Brown clusters improved the recall by 5.3% for SemEval data and 5.1% for the i2b2 data, respectively. This is consistent with the previous research from the biomedical literature\(^\text{14}\). The performances of using the two word embeddings are comparable between the two challenge data sets. The ranking based word embedding algorithm from Collobert et al.\(^\text{23}\) performed slightly better than the word2vec.

Among the three strategies for using the neural word embeddings, the binarized embedding feature method achieved the best F1-score on both of the challenge data sets. Previous research\(^\text{14}\) from the biomedical literature showed that when using a moderate sized corpus (about 42,402 sentences), the Brown cluster feature is superior to the raw word embedding feature (1.2% vs 0.49% on JNLPBA corpus, and 2.1% vs 1.53% on BioCreAtIvE II GM corpus). Our study showed that the binarized embedding feature derived from a much larger corpus (403,871 notes) could be better than the Brown cluster feature (2.3% vs 1.7 for the i2b2 data and 2.4% vs 1.3% for the SemEval data). This could be explained on the basis of the study by Wang and Manning\(^\text{25}\), where the authors showed that the discrete high-dimension feature space works better in conventional machine learning models. Another possible reason may be that the marginal benefit of capturing the multi-aspect relations from a big unlabeled corpus is higher than the benefit from a moderate sized corpus. The distributed prototype feature was comparable to the binarized embedding feature. However, the distributed prototype feature benefitted the precision more than the binarized embedding feature.

To examine what the neural word embeddings captured in the real-valued matrix, we calculated the nearest neighbors using the embedding. Table 4 shows several nearest neighbor examples from the ranking based neural word embeddings. We can see that the neural word embeddings capture a wide range of semantic relations in both the general English domain (e.g., number, time unit, verb) and the clinical domain (e.g., gender, modifier, disorder, laterality, body location, medication). The embeddings also captured the semantic relations involving the clinical abbreviations (e.g., yr-year, l-left and r-right).
Table 4. Examples of nearest neighbor words from the ranking based word embeddings.

<table>
<thead>
<tr>
<th>Word</th>
<th>Top ten nearest neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>two three four several five six another long 0 large</td>
</tr>
<tr>
<td>year</td>
<td>week month weeks yr years days yo old months wk</td>
</tr>
<tr>
<td>stopped</td>
<td>restarted initiated discontinued held started begun added weaned titrated diuresed</td>
</tr>
<tr>
<td>female</td>
<td>male woman man gentleman ga gestation boy m infant</td>
</tr>
<tr>
<td>mildly</td>
<td>moderately markedly slightly severely grossly diffusely somewhat widely relatively extremely</td>
</tr>
<tr>
<td>enlarged</td>
<td>prominent edematous thickened widened collapsed dilated opacified atrophic calcified imaged</td>
</tr>
<tr>
<td>right</td>
<td>left l r bilateral anterior rt proximal posterior lower upper</td>
</tr>
<tr>
<td>atrium</td>
<td>ventricle subclavian arm calf forearm jugular thigh orbit flank elbow</td>
</tr>
<tr>
<td>warfarin</td>
<td>allopurinol decadron methadone labetalol hydrochlorothiazide captopril spironolactone fluconazole haldol metformin</td>
</tr>
</tbody>
</table>

This research has limitations. The corpora used in this research are well preprocessed. However, the real world clinical notes are much noisy and heterogeneous. The comparison between neural word embeddings was conducted on a single conventional machine learning model (CRFs). The neural network model usually has high complexity, which may not fit well on the small corpus. We didn’t consider the hybrid systems and the existing knowledge bases (such as UMLS). We simplified the system to examine the promise of neural word embeddings – automatically learn knowledge from unlabeled clinical corpora. The best models in the i2b2 challenges (with the best F1-score of 0.852) and SemEval challenge (with the best F1-score of 0.813) show that the existing knowledge and more complex feature combinations could further improve the performance. It’s interesting to further examine the world embeddings in the deep neural network based classifiers and combine the existing knowledge bases into word embeddings.

Automatic feature learning from deep neural networks explore rich feature spaces, thus saving the clinical NLP researchers from time-consuming feature engineering. This study showed promising results of using deep neural networks to capture distributed word representations from a large unlabeled clinical corpus to improve the performance of clinical NER systems. Compared with labeled corpus, the unlabeled corpora are much easy to collect. The distributed word embedding features can be adapted to any other clinical NLP system. Moreover, this unsupervised knowledge is low cost - without any involvement of domain knowledge.

Conclusion

This paper studied the neural word embeddings for clinical NER. We systematically compared two popular neural word embedding algorithms and three strategies for deriving distributed word representation features from word embeddings. We also compared the distributed word representation feature with another widely used Brown cluster feature. Evaluation using two challenge datasets showed that the binarized embedding features derived from a large unlabeled corpus could remarkably benefit the clinical NER systems. The word embedding features can be easily adapted to any other clinical NLP research.

Acknowledgement

This study was supported by grants from the NLM 2R01LM010681-05, NIGMS 1R01GM103859 and 1R01GM102282. We would like to thank the 2010 i2b2/VA challenge organizers and the 2014 SemEval challenge organizers for the development of the corpora used in this study.

References

Citation Sentiment Analysis in Clinical Trial Papers

Jun Xu, Ph.D., Yaoyun Zhang, Ph.D., Yonghui Wu, Ph.D., Jingqi Wang, M.S., Xiao Dong, M.D., Hua Xu, Ph.D.
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

ABSTRACT

In scientific writing, positive credits and negative criticisms can often be seen in the text mentioning the cited papers, providing useful information about whether a study can be reproduced or not. In this study, we focus on citation sentiment analysis, which aims to determine the sentiment polarity that the citation context carries towards the cited paper. A citation sentiment corpus was annotated first on clinical trial papers. The effectiveness of n-gram and sentiment lexicon features, and problem-specified structure features for citation sentiment analysis were then examined using the annotated corpus. The combined features from the word n-grams, the sentiment lexicons and the structure information achieved the highest Micro F-score of 0.860 and Macro-F score of 0.719, indicating that it is feasible to use machine learning methods for citation sentiment analysis in biomedical publications. A comprehensive comparison between citation sentiment analysis of clinical trial papers and other general domains were conducted, which additionally highlights the unique challenges within this domain.

INTRODUCTION

Sentiment analysis aims to determine the sentiment polarity conveyed through a segment of text with respect to a specific entity (opinion target)\textsuperscript{1,2}. This entity can be individuals, objects, or topics. With the expanding use of social media, such as blogs and social networks starting from the year 2000, sentiment analysis has attracted increasing attention. Massive opinion data conveying personal sentiments are available online, which were very valuable for business intelligence applications, such as the product search engines\textsuperscript{3}. These intelligent systems are supported by sentiment analysis methods that automatically identify the sentiments from product and service reviews\textsuperscript{3,4}. A quick perusal of existing work shows that most of the state-of-the-art studies approach sentiment analysis as a classification problem. Depending on the granularity of the text regions, sentiment analysis can be conducted at three levels, i.e., document-level\textsuperscript{5}, sentence-level\textsuperscript{6}, and aspect-level\textsuperscript{7}. Typically, the sentiment corpora are composed of reviews, e.g. movie reviews\textsuperscript{5} and product reviews\textsuperscript{6}, news articles\textsuperscript{8,7} and tweets\textsuperscript{8,9}. Researchers have applied different strategies for sentiment analysis, including keyword spotting, lexical affinity, statistical methods, and concept-level techniques\textsuperscript{10}. In the biomedical domain, sentiment analysis has already been used for detection of adverse drug reactions\textsuperscript{11}, assessment of suicide risk\textsuperscript{12}, evaluation of doctor service\textsuperscript{13}, hospital quality survey\textsuperscript{14}, and public health policy making\textsuperscript{15}.

Citation sentiment analysis is an application of sentiment analysis in citation content analysis\textsuperscript{16}, which aims to determine the sentiment polarity that the citation context carries towards the cited paper. In citation sentiment analysis, the opinion target is the cited paper, and the categories of sentiment polarity could be either positive, negative or neutral. Citation content analysis studies are useful for investigating the numerical, literal, and sociocultural\textsuperscript{17} aspects of citations, focusing on the classification of citation functions\textsuperscript{18}, roles\textsuperscript{19}, and the academic influence measures\textsuperscript{20}. Some of the categories in these studies are also related to citation sentiments, e.g., ‘Contrast/Conflict’ and ‘Similarity/Consistency’ in the study by Agarwal et al\textsuperscript{19}. Of late, citation sentiments analysis has emerged as a novel research topic in this area. Athar and Teufel\textsuperscript{21,22} proposed to use citation sentiment analysis to identify whether a particular act of citing was done with positive or negative intentions in computational linguistic publications.

One potential application of citation sentiment analysis is to improve bibliometric measures. In traditional bibliometrics, the impact of a research article is usually measured by citation frequency. Diverse citation counting-based measures, like the Hirsch index\textsuperscript{23} and the g-index\textsuperscript{24}, offer a quantitative proxy to determine the impact. However, all citations are treated equally in these measures, without making any distinctions between the essential function of a citation, i.e., whether the cited work is credited or criticized. Citation sentiment analysis could therefore be used to improve the existing bibliometrics measures by introducing different weights, based on the
specific sentiments of the citations. Moreover, it can provide evidence for scientific authoring support\textsuperscript{25} and citation bias analysis\textsuperscript{26,27}.

Another potential application of citation sentiment analysis in the biomedical domain is to detect non-reproducible studies. Over the last few years, the reproducibility of existing research has become an important issue in the biomedical domain. In a recent research paper, the researchers from Amgen Inc. found that 47 of 53 ‘landmark’ oncology publications could not be replicated \textsuperscript{28}. Moreover, increasing numbers of research articles are being retracted from prestigious journals due to the reproducibility issue. Therefore, it is important to detect non-reproducible studies as soon as possible, to avoid wasting resources, e.g., expensive drug-discovery projects that attempt to confirm non-reproducible results. Here we propose that the sentiment polarity conveyed in the citation context could be a useful clue denoting the reproducibility of the cited paper. For example,

\textit{I. Thus, inhibition of SATB1 expression does not appear to alter the aggressive phenotype of breast cancer cell lines in vitro, in contrast to the results reported by Han et al. [1].}

In the above example I, a negative opinion is expressed towards the cited paper (reference id: 1), implying that the citing work could not reproduce the cited work. The reproducibility of a study could thus be evaluated by aggregating the sentiments dispersed in the publications citing it as a reference. Thus, citation sentiment analysis comprises the preliminary step towards reproducibility evaluation. However, there are limited studies about citation sentiment analysis in the biomedical domain. The most related study is from Yu\textsuperscript{25}, where the author analyzed the needs of citation sentiment analysis and suggested that it should consider the sentiment of both specific claims and the citations.

To address the critical problem of evaluation of research reproducibility, this study proposed to identify citation sentiment in clinical trial papers using machine learning methods. A citation-level annotated sentiment analysis corpus composed of 285 clinical trial related publications was first constructed. Based on the corpus, we developed a machine learning based classifier and systematically compared different features and feature combinations. The experimental results show that it is feasible to use machine learning classifiers to identify citation sentiments from biomedical publications. One thing to be noted is that both the annotation and detection of citation sentiment are context-enhanced. To the best of our knowledge, this is one of the first studies that addresses automated citation sentiment analysis in the biomedical domain.

METHODS

This study attempts to classify the sentiment polarity of a citation in clinical trial papers, i.e., the sentiment polarity relation between the citing and the cited paper, based on the sentiment analysis of the content of citation context. We first annotated a citation sentiment analysis corpus, which contains discussion sections extracted from 285 clinical trial papers. The citation sentiment polarity was annotated at the citation-level by following an annotation guideline. A simple rule-based method was used to extract the citation context, which is a set of on-topic sentences. The citation sentiment polarity was then classified using machine learning methods incorporating features extracted from the citation context. The performance of citation sentiment analysis was evaluated using the 10-fold cross-validation method.

Data Set Annotation

Data Set Selection The data we used in this study are derived from clinical trial papers. Some clinical trials are ‘self-correcting’ studies, which are founded on the replication of the earlier trials. The clinical trial paper contains the opinionated citations we wanted and satisfied the purpose of our concern on the reproducibility of biomedical research. We used the query “clinical trial[Title] AND ("research and review articles"[filter] AND "open access"[filter] AND "2004/06/01"[PubDate] : "2014/05/31"[PubDate])” to search the Pubmed Central (PMC)\textsuperscript{*}. Then we downloaded all the papers in the search results from the PMC archive using OAI-PMH service\textsuperscript{†}. We randomly selected 285 papers for corpus construction. The clinical trial paper in the biomedical domain has a purely descriptive style and a very standardized structure known as Introduction, Methods, Results, and Discussion (IMRAD). Most of the opinionated citations were found in the discussion section. Therefore, we only extracted the content of the discussion section. Multiple-references citation, e.g. ‘[1], [3]’ or ‘[3-10]’, was treated as one citation, since they share same citation sentiment polarity. Based on the above criteria, we obtained 4182 citations.

\textsuperscript{*} http://www.ncbi.nlm.nih.gov/pmc/
\textsuperscript{†} http://www.ncbi.nlm.nih.gov/pmc/tools/oai/
**Annotation Schema** In previous work\textsuperscript{21, 22}, each sentence which has a citation was treated as the citation context, and the sentiment was labeled at the sentence level. In our study, we annotated each citation based on its context. The citation level annotation is applied for the following reasons. First, in science papers, the context scope of a citation varies widely, from a clause to several sentences, even a paragraph. Second, there are some sentences which have more than one citation and have different sentiment polarities.

Based on our observation of the problem, we developed a simple and practical annotation scheme. We illustrate it as a decision tree, as shown in Figure 1 below. There are no instructions about the use of cue phrases. This study does not consider the correctness or incorrectness of the claims, ideas, results and conclusions presented in the citing or cited paper. The opinion target is the cited paper, not a scientific claim. For example, both citations in “Xxx et al. \textsuperscript{[3]} concluded that ..., which did not agree with Yyy et al. \textsuperscript{[7]}.” are labeled as neutral.

**Annotation Process** For the annotation task, three annotators were employed. According to the scheme discussed earlier, all citations were annotated by the two primary annotators. One of the annotators was a physician and served as the domain expert. The third annotator was involved when the two primary annotators have differing annotations for the same citation.

![Decision Tree](https://example.com/decision_tree.png)

**Figure 1. Simple annotation scheme designed as a decision tree**

**Citation Context Extraction**

Citation context is defined as the textual statement that contains the citation. It is a set of on-topic (describing the target citation) sentences surrounding the citation. In our study, we hypothesized that citation context will provide relevant and related information which will help identify the citation sentiment. Most of the citations’ sentiment can be determined by the sentence which contains the citation itself. However, some of them need sentential context. A simple rule-based method was developed to extract the context content for each citation. First, for each citation (target citation), we extracted the whole sentence that contains it. Then, if the next sentence did not contain other citation(s), and there was a contrastive discourse relation between the two sentences, the next sentence was also added to the citation context. To determine the contrastive relation, a small set of cue-phrase-based patterns were used, including ‘however (.)...’, ‘conversely, ...’, etc. The citation context extraction partially solves the problem of complex discourse structure, which may have expression of true opinion.

**Citation Sentiment Polarity Classification**

**Features**

Considering the characteristics of citation sentiment analysis, we systematically extracted the following features, including word n-grams, sentiment lexicons and problem-specified structure features.
(1) Word n-grams

All unigrams, bigrams and trigrams in the citation context were extracted as features. Since negation words always affect the polarity, they have to be taken into consideration in sentiment analysis. The total count of negation phrases found within the citation context was therefore extracted as features. Furthermore, artificial words like 'NOT_w' were extracted if the word ‘w’ was in the scope of negation. In this study, the negation scope was set as 2-words to the right of negations.

(2) Sentiment lexicons

Researchers use sentiment words to express opinions, which could be used as clues for citation sentiment analysis. We manually created a biomedical research paper specific sentiment lexicon. The lexicon includes 53 positive and 46 negative phrases. They were extracted from the corpus we built. Most of them are high frequent words found in biomedical literature. The presence and absence of a sentiment word were also extracted as features in addition to their sentiment polarities. We also assigned opposite polarity if the sentiment word was detected within the scope of a negation word.

(3) Structure features

Based on the observation of the examples, we found that the sentiment of citation was constrained by the intra- and inter-sentence features, including contrast discourse relation, sentiment words, negation words, comparative relation, related position and direction information of citing work, target, and other non-target cited work. We extracted and combined all these features in an order to generate structure features, which have better representation power by retaining intra- and inter-sequential relations. In scientific writing, researchers usually use a number of particular ways to express their sentiment regarding a citation, such as comparative expression and complex discourses. Thus, we designed a feature extraction pipeline to capture those complex structures. As shown in Table 1, the feature extraction pipeline composed of two steps: 1) Extract and tag all of the citing work tokens, comparative tokens, negation words, sentiment tokens, contrastive discourse cue words in the citation context. This is rule-based processing. We built a series of cue-phrase-based patterns to recognize concept entities like citing works, negation words, comparative expressions, and contrast discourse relations; 2) Generate the unigram, bigram and trigram of recognized concepts as features. If there is a contrast discourse relation in the citation context, we also generated the direction feature. It represented the position and direction information of citing, cited works and the contrast cues, as shown in the last row of Table 1.

Table 1. An illustration of the structure features extraction.

<table>
<thead>
<tr>
<th>Example</th>
<th>Our data are generally consistent with that of other studies [TC], but not with studies where a single dose of paracetamol was administered [OC].</th>
<th>These values are lower than those reported by French et al. [20].</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>[Our data] CITINGWORK are generally [consistent with] POSITIVE that of other studies [TC] [OC], [but] CONTRAST [not] NEGATION with studies where a single dose of paracetamol was administered [OC] [OC]</td>
<td>[These values] CITINGWORK are [lower] COMPARATIVE [than] THAN those reported by French et al. [TC] [OC]</td>
</tr>
<tr>
<td>Direction</td>
<td>‘CITINGWORK_CONTRAST_DIR’, ‘TC_CONTRAST_DIR’, ‘CONTRAST_OC_DIR’</td>
<td></td>
</tr>
</tbody>
</table>
Machine Learning Algorithms

Before the feature extraction, the reference markers in the sentence, such as ‘[3]’, ‘[8, 9]’, and ‘[2, 6-10]’ etc., were identified. A citation context may contain several citations. Consequently, we replaced the text of target citation and other citations with the tokens ‘[TC]’ and ‘[OC]’, respectively. Then, we extracted the features from the citation context. Each citation context was represented as a feature vector. In this study, Support Vector Machines (SVMs) was employed for its known performance on previous sentiment analysis tasks. More specifically, the implementation of LIBLINEAR was used in our study. Parameters for SVM classifiers were optimized based on training set.

Evaluation

This study adopted the standard evaluation metrics. We reported the accuracy (Acc.), Micro- and Macro- F across all the categories. As the corpus is imbalanced, we further reported the precision (P), recall (R) and F-measure (F) for each category. In the experiments, 10-fold cross validation was performed. All the performances given below are the average of 10 folds. We report the results with different combinations of features.

RESULTS

Table 2 shows the descriptive statistics of the corpus. The corpus contains 702 positive, 308 negative and 3172 neutral samples. The KAPPA value among the annotations from the two primary annotators’ was 0.504. It is a moderate agreement level.

Table 2. The statistic information of the corpus.

<table>
<thead>
<tr>
<th>Number of Documents</th>
<th>Number of Citations</th>
<th># Positive</th>
<th>#Neutral</th>
<th>#Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>285</td>
<td>4182</td>
<td>702 (16.79%)</td>
<td>3172 (75.85%)</td>
<td>308 (7.36%)</td>
</tr>
</tbody>
</table>

Table 3 shows the results of the citation sentiment analysis on the corpus we annotated. Different combinations of features were tested. Default parameters of LIBLINEAR package were used. We treat the results achieved using word n-gram features as the baseline. By adding sentiment lexicon features or structure features, the recall as well as the overall performance of both negative and positive class increased significantly. With a combination of all the features, we reached the best overall performance (accuracy of 0.870) and the highest F for each class, with 0.511 for “negative”, 0.924 for “neutral”, and “0.723” for “positive” respectively.

Table 3. Results of 10-fold cross validation using different combination of features.

<table>
<thead>
<tr>
<th>Features</th>
<th>Overall</th>
<th>Per Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>Micro-F</td>
</tr>
<tr>
<td>(1)</td>
<td>0.853</td>
<td>0.834</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)+(2)</td>
<td>0.869</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)+(3)</td>
<td>0.868</td>
<td>0.856</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)+(2)+(3)</td>
<td>0.870</td>
<td>0.860</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Feature sets: (1) Word n-grams; (2) Sentiment lexicons; (3) Structure features.
DISCUSSION

We conducted a study on identifying the sentiment polarity from the citation context in clinical trial papers using machine learning methods. The combined features from the word n-grams, the sentiment lexicons and the structure information achieved the highest Micro-F score of 0.860 and Macro-F score of 0.719, validating the promise of using automated machine learning methods to analyze the sentimental polarities in citation context from biomedical literature.

In this study, we constructed a citation analysis corpus from 285 randomly selected clinical trial related research articles. To the best of our knowledge, this is the first citation sentiment analysis corpus in the biomedical domain. The annotators had a moderate inter-annotator agreement with a Kappa of 0.5040, suggesting that citation sentiment identification is challenging even for domain experts. We also found that biomedical researchers expressed their sentiments in many different ways, which cannot be identified by simply using the sentiment lexicons. Sometimes, the annotators had to link distant citation context to figure out the sentiment polarity, which made it more difficult to automatically identify the sentiment using machine learning methods.

Compared to general sentiment analysis study on product reviews, news articles, and Twitter, citation sentiment analysis on biomedical literature presented several unique challenges.

(1) Most of the citations were neutral with respect to sentiment. The purpose of citing was to gain information, such as background, methods and data information from the cited work. Authors typically used objective words in the citation context. As a result, the corpus we constructed for citation sentiment analysis is imbalanced, as shown in Table 2. An example sentence is shown below:

II. CU has been widely studied throughout literature for its anti-inflammatory [13, 14], anti-oxidant [15], antibacterial [16] and wound healing [17], properties.

(2) Sentiment polarity was often expressed as a comparative sentence. Typically, the authors used comparative claims to express opinions, either refuting or confirming a previous work, as comparative claims clearly describe similarity or difference relation between the citing and the cited work. The negativity or positivity was implied by the comparison with the authors’ work. An example is noted below:

III. Compared with previous studies using intratumoral or intravenous infusion of rAd-p53 [6], this study achieved higher overall clinical response rates with intra-arterial administration.

(3) Scientists are reluctant to be critical when they cannot reproduce previous results. In some cases, the express of citation sentiment is very implicit without any explicit linguistic cues, such as negative words. Please see below for an example.

IV. It has been reported that the inclusion of micronutrients can decrease the number of acute respiratory infections [36]. In the present study, no differences were found in respiratory tract infections among treatment groups.

(4) The context scope of a citation varied widely, from a single clause to multiple sentences. Meanwhile, adversative transition words were used extensively to connect these clause or sentences. The citation sentiment was constrained and implied by these intra- and inter-sentence discourse relation. A multiple sentences example is provided below.

V. Previous reports by Ulrichs et al. have shown that IFN-g producing cells against ESAT6 in tuberculosis patients increase post-tuberculosis therapy [18]. However, we were unable to demonstrate any difference in ESAT6-induced IFN-γ responses between patients prior to and post-treatment. This is in concordance with a recent report by Coussens et al. [14] that did not show any change in ESAT-6 induced IFN-g in patients’ post-antituberculous therapy.

In conclusion, imbalanced data, comparative expression, implicit sentiment and complex discourse structure made the citation sentiment analysis more challenging in the biomedical domain.

The clinical trial related biomedical literature is usually about biomedical findings. The biomedical researchers express their sentiments indicating whether his work confirms, supports, or agrees with the cited paper. The literature in other domains, such as the computer science, is more about theoretical models, algorithms and application systems. Therefore, the scientists used different ways to express sentiments in different domains. For
examples, “agree with” and “confirm” were always used within a positive citation context in the biomedical domain. However, they were uncommon in the literature of computer science.

Although the overall performance on the three sentiment categories is promising, the performance for the negative category still needs to be improved. We conducted an error analysis for the negative category and found that the low performance was caused by the implicit sentiment expression and complex discourse structures. Normally, the authors tried their best to avoid giving explicit personal effect especially negative credit to the cited paper. Further, the intentions of the authors were not sometimes available to the content analyst. For example, in the sentence “We did not observe any reductions in mortality, as did Ferrer and colleagues [14].”, more subjective interpretation to assign citations to the defined sentiment categories is required. Textual entailment techniques may be required to determine whether the mentioned statement contradicts the claim of the cited paper or not. In our study, for simplicity, we only proposed a rule based citation context extraction method. The on-topic sentence could potentially be missed if the cue words are out the scope of the two continuous sentences. Furthermore, other discourse relations related to sentiment analysis were not taken into account.

The method we used is not proposed for learning from imbalanced data. Some state-of-the-art methods for imbalanced learning may improve the performance of minority classes. That’s one of the aims of our future work. For minority classes, it is hard to collect enough samples for training. However, massive citations can be obtained easily. Semi-supervised methods could be a potential way to handle the imbalanced data problem for citation sentiment analysis. Another limitation of this study is that data were collected from clinical trial papers only; therefore more optimizations would be required when we extend the work to all biomedical papers.

CONCLUSION

In this study, we manually annotated a citation sentiment corpus for clinical trial papers and developed a machine learning approach to determine sentiment polarity of biomedical citations. We conducted experiments to examine the effectiveness of the common features (i.e., word n-gram and sentiment lexicons) and structure features in classifying citation sentiment. Our results and analysis indicate that it is feasible to determine biomedical citation sentiment using machine learning approaches, although challenges still exist. In the future, we will enhance our approach by employing methods for imbalanced data and discourse relations analysis techniques that could extract more informative context features.

Acknowledgments

This study is supported in part by grants from NLM 2R01LM010681-05, NIGMS 1R01GM103859 and 1R01GM102282. The first author (JX) is partially supported by NSFC 61203378.

References

tcTKB: an integrated cardiovascular toxicity knowledge base for targeted cancer drugs

Rong Xu¹, and QuanQiu Wang²

¹Department of Epidemiology and Biostatistics, Institute of Computational Biology, School of Medicine, Case Western Reserve University, Cleveland OH 44106
²ThinTek, LLC, Palo Alto, CA 94306

Abstract

Targeted cancer drugs are often associated with unexpectedly high cardiovascular (CV) adverse events. Systematic approaches to studying CV events associated with targeted anticancer drugs have high potential for elucidating the complex pathways underlying targeted anti-cancer drugs. In this study, we built tcTKB, a comprehensive CV toxicity knowledge base for targeted cancer drugs, by extracting drug-CV pairs from five large-scale and complementary data sources. The data sources include FDA drug labels (44,979 labels), the FDA Adverse Event Reporting System (FAERS) (4,285,097 records), the Canada Vigilance Adverse Reaction Online Database (CVAROD) (1,107,752 records), published biomedical literature (21,354,075 records), and published full-text articles from the Journal of Oncology (JCO) (13,855 articles). tcTKB contains 14,351 drug-CV pairs for 45 targeted anticancer drugs and 1,842 CV events. We demonstrate that CV events positively correlate with drug target genes and drug metabolism genes, demonstrating that tcTKB in combination with other data resources, could facilitate our understanding of targeted anticancer drugs and their associated CV toxicities.

Introduction

Treatment outcomes in cancer patients have dramatically improved since the introduction of targeted drugs. However, targeted drugs are often associated with unexpectedly high cardiovascular (CV) toxicities in cancer patients [1-2]. The mechanisms by which targeted drugs exert their toxic effects on heart and vasculature in cancer patients are not well-understood [3-4]. To ensure safe personalized cancer treatment, research efforts are needed to understand CV toxicities associated with targeted drugs. Systematic and integrated approaches to studying CV events associated with targeted drugs have high potential for elucidating the complex pathways underlying anti-cancer drugs, identifying the on- and off-targets of undesirable CV events, and predicting unknown CV toxicities [5-7]. However, systematic study of targeted drug-induced CV toxicities has been hampered by the lack of a comprehensive and machine-understandable knowledge base of drug-CV associations. The relevant knowledge is instead buried throughout multiple disparate and complementary information sources in varying formats. It was recently demonstrated that 39% of serious events associated with targeted cancer drugs were not reported in clinical trials and 49% were not described in FDA drug labels [8]. Therefore, in order to build a comprehensive knowledge base of CV toxicities associated with targeted drugs, it is important to extract knowledge from multiple complementary data sources.

Recently, we extracted targeted anticancer drug-associated CV events from the U.S FDA Adverse Event Reporting System (FAERS) (4,285,097 records) [9]. We also developed text classification, relationship extraction, signaling filtering, and signal prioritization algorithms to extract targeted anticancer drugs associated side effects, including CV events from 13,855 full-text articles and embedded tables from the Journal of Oncology (JCO) published between 1983 and 2013 [10-11]. In this study, we built tcTKB (the CardioToxicity Knowledge Base for Targeted Cancer Drugs) by combining drug-CV pairs extracted from FAERS and JCO articles with pairs from another three large-scale and publicly available datasets: U.S. Food and Drug Administration (FDA) drug labels (44,979 drug labels), the Canada Vigilance Adverse Reaction Online Database (CVAROD) (1,107,752 records), and the vast corpus of published biomedical literature (21,354,075 MEDLINE records).

Drug toxicity knowledge contained in these five data sources is largely complementary. FDA drug labels contain known adverse events associated with commercial drugs, which are mainly gleaned from controlled clinical trials. Drug-CV pairs from FDA drug labels are highly accurate (high precision), however the recall may be limited since
they are mainly obtained from pre-marketing clinical trials for well-controlled patients (i.e. patients with less co-morbidities or younger patients). The two post-marketing surveillance systems (FAERS and CVAROD) contain both voluntary and mandatory reports of suspected drug adverse events from health-care professionals, consumers, and pharmaceutical companies for drugs used in less controlled 'real-world' patient populations. FAERS is the main spontaneous reporting system overseen by the U.S. FDA. Mining drug-side effect (drug-SE) relationships from FAERS is a highly active research area. Data mining algorithms such as disproportionality analysis, correlation analysis, and multivariate regression have been developed to detect adverse drug signals from FAERS [12-14]. Recently, we developed signaling extraction, prioritization, filtering algorithms and extracted a total of 11,173 drugCV pairs, representing 39 targeted cancer drugs and 1095 CVs, from FAERS [9]. In this study, we will extract drug-CV pairs from CVAROD, the main spontaneous reporting system overseen by Health Canada. CVAROD contains more than one million patient records, however, research effort in mining drug safety signals from CVAROD is significantly less compared to efforts in mining FAERS. Drug-CV pairs extracted from these two post-marketing surveillance systems are comprised of known true positives (those included in FDA drug labels), unknown true positives, and false positives. The main challenge is to differentiate between unknown true positives and false positives.

JCO is the official journal of the American Society of Clinical Oncology and the leading journal in oncology. JCO articles not only include pivotal clinical trials that have led to drug approval, but also trials that are still in investigational stages and even failed trials. In one of our recently studies, we downloaded a total of 13,855 full-text JCO articles published between 1983 and 2013. We combined automatic table classification and relationship extraction approaches to extract anticancer drug-associated side effects from a total of 31,255 tables embedded in these JCO articles [10]. We also developed an integrated system combining text classification, relationship extraction, signal filtering, and signal prioritization algorithms to extract targeted anticancer drug-associated side effects from the full-text part of JCO articles [11]. We demonstrated in our previous studies and in this study that full-text oncological articles contains much drug-associated side effects including CV events that are not captured in FDA drug labels, MEDLINE abstracts, or post-market drug safety surveillance systems.

Currently, more than 22 million biomedical records are publicly available on MEDLINE, making it a rich source of CV toxicity information for drugs at all clinical stages, including drugs in pre-marketing clinical trials, post-marketing clinical case reports and clinical trials. The major challenge in extracting drug-CV pairs from this rich source is that it is buried in free-text format. We recently develop approaches to extract drug-SE pairs from MEDLINE [15-16]. In this study, we will apply these approaches to extract target anticancer drug-associated CV events from MEDLINE abstracts. In this study, we show that as much as 96% and 67% targeted anticancer drug-associated CV pairs extracted from MEDLINE are not included in FDA drug labels and FAERS, respectively, demonstrating the need of building tcTKB from multiple complementary data resources. To the best of our knowledge, this is the first research effort in building a comprehensive CV toxicity knowledge base for targeted drugs from multiple complementary data sources. We demonstrated that this unique knowledge base tcTKB, in combination with other data resources such as drug target databases or drug pharmacogenetics and genomics databases, can facilitate our deeper understanding of the molecular mechanisms underlying the unexpectedly high incidence of CV toxicities associated with many targeted anticancer drugs.

Data and methods

The data resources and methods that were used to construct tcTKB is depicted in Figure 1 and described later.

Data

The four data sources for extracting targeted anticancer drug-associated CV pairs are summarized in Table 1 and described in the sections that follow.

FDA drug labels We downloaded a total of 44,979 drug labels, including 21,610 human prescription labels and 23,369 human OTC labels from DailyMed. DailyMed is maintained by the National Library of Medicine (NLM)
and provides high quality FDA package inserts information about marketed drugs. The majority of drug side effect information captured on FDA drug labels is obtained from clinical trials, while some is obtained from post-marketing surveillance. We used the publicly available information retrieval library Lucene\(^2\) to create a local FDA drug label search engine with indices created on drugs, section headers such as “Indications,” “Contraindications,” and “Adverse Reactions;” and sentences. Each sentence was associated with a drug and a subsection header name.

**The FDA Adverse Event Reporting System (FAERS)** FAERS is the prominent post-marketing drug safety surveillance system maintained by the U.S. FDA. In our recent study, we downloaded a total of 4,285,097 records from FAERS for the time period of 2004 to 2012\(^3\). We extracted a total of 11,173 drug-CV pairs, representing 39 targeted cancer drugs and 1095 CVs, from FAERS [9].

**The Canada Vigilance Adverse Reaction Online Database (CVAROD)** CVAROD is the main post-marketing drug safety surveillance system in Canada. A total of 1,107,752 patient records were downloaded\(^4\). File “reactions.txt” contains the reported adverse events. File “drug_product Ingredients.txt” contains drug information. Similar to pairs extracted from FAERS, pairs in CVAROD could contain spurious pairs as well as unknown true positives.

**JCO full text articles** In our previous study, we downloaded a total of 13,855 JCO full text JCO articles published from 1983 through 2013 and extracted anticancer drug-SE pairs from both the text and the tables in the articles [10-11]. We extracted a total of 3,515 drug-CV pairs, representing 45 targeted anticancer drugs and 449 CV events.

**MEDLINE data and local MEDLINE search engine** We downloaded a total of 21,354,075 MEDLINE records (119,085,682 sentences) published between 1965 and 2012 from NLM (http://nlm.nih.gov/Download/index.shtml). Each sentence was syntactically parsed with Stanford Parser [16] using the Amazon Cloud computing service (a total of 3,500 instance-hours with High-CPU Extra Large Instance were used). We created a local MEDLINE search engine

\(^2\)http://lucene.apache.org
\(^3\)http://www.fda.gov/Drugs/
\(^4\)http://www hc-sc gc ca/dhp mps/medeff/databaseon/extract_extrait eng.php

---

Figure 1: The data resources and methods used in constructing tcTKB.

Table 1: Summary of the five data sources of drug-CV associations.

<table>
<thead>
<tr>
<th>Source</th>
<th>Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDA drug labels</td>
<td>44,979 drug labels</td>
<td>Pre-marketing (main) and Post-marketing</td>
</tr>
<tr>
<td>FAERS</td>
<td>4,285,097 records</td>
<td>Post-marketing (U.S)</td>
</tr>
<tr>
<td>CVAROD</td>
<td>1,107,752 records</td>
<td>Post-marketing (Canada)</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>21,354,075 records</td>
<td>All stages</td>
</tr>
<tr>
<td>JCO</td>
<td>13,855 full-text articles</td>
<td>pre-marketing</td>
</tr>
</tbody>
</table>
with indices created on sentences, their corresponding parse trees, and abstracts.

**Lexicon of targeted cancer drugs** The 45 targeted cancer drugs was obtained from the National Cancer Institute\(^5\).

**Lexicon of cardiovascular event (CV) terms** We built a lexicon of CV terms based on MedDRA, a medical terminology widely used for classifying adverse events associated with drugs and other medical products\(^6\). The adverse events captured on FDA drug labels, FAERS, and CVAROD are coded with MedDRA terms. We created a lexicon of CV terms by finding all leaf nodes with the ancestor “vascular disorders” or “cardiac disorders.” This lexicon consisted of a total of 1,712 CV terms, including 1,269 vascular disorders and 527 cardiac disorders. In order to capture all the term variations in MEDLINE, we expanded the CV lexicon by including the synonyms of the terms in the lexicon. The term-synonym mappings were derived from UMLS Semantic Network [17]. After expansion, the CV lexicon consisted of 27,547 terms.

**Methods**

**Extract drug-CV pairs from FDA drug labels** We used each of the 45 * 1,712 drug-CV combinations (45 targeted drugs and 1,712 CV terms) as search queries to the local FDA drug label search engine. Drug-CV pairs that appeared in sentences with the header “Adverse Reactions” were retrieved. We extracted a total of 259 drug-CV pairs from FDA drug labels, representing 15 targeted cancer drugs and 75 CV events. These pairs are of high quality and represent known cardiovascular events associated with targeted cancer drugs derived from both pre-marketing clinical trials (major) and post-marketing surveillance (minor).

**Extract drug-CV pairs from FAERS** Drug-CV pair extraction from FAERS was done in our previous study [9]. After file linking, drug entity recognition and mapping, and CV entity recognition, we obtained a total of 11,173 drug-CV pairs, representing 39 (out of 45) targeted drugs and 1,095 (out of 1,712) CV events.

**Extract drug-CV pairs from JCO articles** Drug-CV pair extractions from JCO articles (including both full-text and embedded tables) were done in our previous study [10-11]. We extracted a total of 3,515 drug-CV pairs, representing 45 (out of 45) targeted anticancer drugs and 449 (out of 1,712) CV events.

**Extract and prioritize drug-CV pairs from CVAROD** We extracted drug-CV pairs from CVAROD by linking three downloaded files: (1) Report_Drug.txt, which provides report drug identifiers and adverse reaction report numbers; (2) Drug_Product_Ingredient, which provides drug identifiers and active ingredient names; and (3) Reactions.txt, which provides adverse reaction report numbers and adverse reaction terms. We first linked the file “Drug_Product_Ingredient” with file “Reactions.txt” using the identifiers specified in “Report_Drug.txt”. We then extracted drug-CV pairs from the linked file. Unlike drug strings in FAERS, drugs in CVAROD were already mapped to their active ingredients, therefore no additional concept recognition and mappings were necessary. We filtered the drugs with the lexicon of 45 targeted cancer drugs. The same named entity recognition for CV terms as was completed for FAERS [9] was also performed for the data collection from CVAROD. In total, we obtained 1,160 drug-CV pairs, representing 31 targeted cancer drugs and 300 CVs.

While drug-CV pairs extracted from FDA drug labels are known true positives, pairs extracted from CVAROD contain known true positives, unknown true positives, and true negatives. In order to prioritize drug-CV pairs extracted from CVAROD, we implemented and compared six ranking algorithms in prioritizing true signals, including ranking by pairs’ frequency counts (FREQ) in FAERS, and five commonly used Disproportionality Analysis (DPA) statistical signal detection approaches: relative reporting ratio (RRR), proportional reporting ratio (PRR), reporting odds ratio (ROR), phi coefficient (PhiCorr), and information component (IC). The five DPAs are currently the most widely used approaches for automated signal detection in FAERS [11]. All these DPA methods are based on frequency analysis of 2x2 contingency tables to estimate statistical association between drugs and SEs and it intends to quantify the degree to which a drug-SE pair co-occurs disproportionally in the database. These five DPA methods differ by the statistical adjustments they apply to account for low counts.

\(^{5}\)http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted

\(^{6}\)http://www.meddrampsso.com/
In order to compare different ranking methods, we used 11-point interpolated average precision, a commonly used measure in evaluating information retrieval results [18]. For each ranked list, the interpolated precision was measured at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0. At each recall level, we calculated the arithmetic mean of the interpolated precision. A composite precision-recall curve showing 11 points was then graphed. In order to compare these six ranking approaches in prioritizing true signals, we used the 259 drug-CV pairs extracted from FDA drug labels as the evaluation dataset, which consisted of known true positives. If a ranking algorithm ranks known true positives highly, we can reason that it also ranks many unknown true positives highly. Note: this evaluation dataset was not used to calculate the true precisions and recalls, but to compare the six ranking approaches in prioritizing true signals.

**Extraction and manual curation of drug-CV pair from MEDLINE** We used each of the 45 * 27,547 drug-CV combinations for the 45 targeted cancer drugs and 27,547 CV terms as search queries to the local MEDLINE search engine. Sentences, their associated parse trees, and abstracts that contained each pair were retrieved. Instead of simply retrieving a pairs co-occurrence counts at both sentence- and abstract-level from the search engine, we added the extra restriction that both drug and CV terms must be noun phrases in retrieved parse trees. This additional restriction was put in place to prevent the extraction of partial drug-CV pairs from sentences. For example, a CV term appeared as a substring in a noun phrase in the sentence where the noun phrase term is not included in the input CV lexicon. We extracted a total of 1,080 drug-CV pairs (38 targeted cancer drugs and 470 CVs) from MEDLINE sentences and 2,469 pairs (40 targeted cancer drugs and 877 CVs) from MEDLINE abstracts.

Unlike drug-CV pairs extracted from FAERS or CVAROD, which are hard to evaluate, pairs extracted from MEDLINE have associated abstracts which can be used for manual curation. We extracted a total of 1,080 drug-CV pairs from MEDLINE sentences and then manually curated these pairs. We used the local MEDLINE search engine to retrieve all the sentences (8,590 in total) wherein the pairs appeared. We then manually classified these 1,080 drug-CV pairs into three classes (CAUSE, TREAT, and NONE) using the sentences (and abstracts when necessary) as evidence. Three curators with graduate degrees in biomedical sciences independently performed the curation. It took an average 25 hours for each curation in annotating these sentences. Majority vote was used to decide the final classification.

**Correlation analysis** We investigated whether drug-drug pairs that shared CV events also tended to share gene targets. We downloaded a total of 10,478 drug-gene pairs from DrugBank [19], a knowledge base for drugs, drug actions, and drug targets. These downloaded drug-gene pairs included a total of 24 targeted cancer drugs. For drug-drug pairs that shared different numbers of CV events, we calculated the average number of shared gene targets.

We investigated whether drug-drugs that shared CV events also tended to share drug metabolism genes. We downloaded a total of 4,399 drug-gene pairs from PharmGKB (the Pharmacogenetics and Pharmacogenomics Knowledge Base) [20], a public repository of genotype and phenotype information relevant to pharmacogenetics. The drug-gene pairs were assigned with subtypes Pharmacokinetics (“PK”) and Pharmacodynamics (“PD”). These downloaded drug-gene pairs included a total of 25 targeted cancer drugs. For drug-drug pairs that shared different numbers of CV events, we calculated the average number of shared metabolism genes.

1 Results

1.1 Description of tcTKB

tcTKB is comprised of a total of 14,351 unique drug-CV pairs extracted from five data sources, representing 45 targeted drugs and 1,842 CV events (Table 2). From the FDA drug labels, we extracted a total of 259 drug-CV pairs, representing 15 targeted drugs and 75 CV events. Unlike pairs extracted from the other three data sources, pairs extracted from FDA drug labels are mainly known true positives. From FAERS, we extracted a total of 11,173 drug-CV pairs, representing 39 targeted cancer drugs and 1,095 CV events. These CV events were reportedly associated with targeted cancer drugs in patients in real-world settings and included many CV events that have not yet been captured in the FDA drug labels. We extracted a total of 1,160 drug-CV pairs from CVAROD for 31 targeted cancer drugs and 300 CV events. This number is 10 times smaller than the number of pairs extracted from FAERS. It is unclear why significantly fewer CV events were reported in CVAROD than in FAERS although the numbers of targeted cancer drugs in both databases are comparable. From MEDLINE sentences, we extracted 1,080 drug-CV pairs for 38 targeted cancer drugs and 479 CV events. From MEDLINE abstracts, we extracted 2,469 drug-CV pairs representing 40 cancer
targeted drugs and 877 CV events. The numbers of drug-CV pairs extracted from MEDLINE were smaller than that from FAERS; however, these pairs contained 40 out of the 45 targeted cancer drugs. A total of 3,515 drug-CV pairs were extracted from JCO articles. These pairs included all 45 targeted drugs, demonstrating that JCO articles is good data resource for targeted anticancer drugs.

### 1.2 Drug-CV pairs extracted from four data sources are largely complementary

We investigated overlaps of drug-CV pairs extracted from FDA drug labels, FAERS, CVAROD, MEDLINE, and JCO. As shown in Table 3, the information in these four sources overlaps, but is largely complementary. Column 2 in Table 3 represents the percentages of drug-CV pairs from FDA drug labels that were also included in the other data sources. Row 2 represents the percentages of drug-CV pairs from the other three data sources that were captured in the FDA drug labels. As shown in both column 2 and row 2, many of the known drug-CV pairs in FDA drug labeling appeared in the other four data sources: 95.8% in FAERS, 46.7% in CVAROD, 38.2% in MEDLINE, and 71.4% in JCO (column 2). However, the opposite is not true. Only very small percentages of drug-CV events reported in other data sources were captured in FDA drug labels (row 2). This low percentage may be due to the following reasons: First, FDA drug labels contain only 15 targeted drugs, while the other data sources contained more targeted drugs. Second, while the drug-CV pairs extracted from FDA drug labels are mostly true positives, the pairs extracted from the other sources may contain false positives and unknown true positives.

Only small percentages of pairs from FAERS were included in other data sources: 2.2% in FDA drug labels, 10.1% in CVAROD, 7.4% in MEDLINE, and 15.3% in JCO (column 3). In addition, a majority of drug-CV pairs from other data sources (except CVAROD) were not included in FAERS (row 3). This indicates that drug-CV pairs contained in FAERS are largely complementary to those in FDA drug labels or MEDLINE. Drug-CV pairs contained in CVAROD are a subset of the pairs from FAERS as 96.7% of the 1,160 pairs also appeared in FAERS. Drug-CV pairs extracted from MEDLINE are largely complementary to those in the other three data sources. For example, only 4.0% of pairs from MEDLINE appeared in FDA drug labels, 33.1% in FAERS, and 12.2% in CVAROD (Column 5). The opposite is also true. Only 38.2% drug-CV pairs from FDA labels, 7.4% from FAERS, and 26.0% from CVAROD also appeared in MEDLINE. It is also evident from the table that full-text articles contain much information not captured in MEDLINE abstracts. For example, only 23% drug-CV pairs from JCO articles also appeared in MEDLINE abstracts.
Figure 2: Ranked precisions at 11 recalls for drug-CV pairs from CVAROD ranked by three ranking measures: frequency (Freq), proportional reporting ratio (PRR) and phi coefficient (PhiCorr). Data for relative reporting ratio (RRR), reporting odds ratio (ROR), and information component (IC) IC are similar to that for PRR and not shown.

### Table 4: Manual curation of drug-CV pairs extracted from MEDLINE sentences.

<table>
<thead>
<tr>
<th>Total</th>
<th>CAUSE</th>
<th>TREAT</th>
<th>NONE</th>
<th>Not in FDA drug labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,080</td>
<td>48.6%</td>
<td>32.9%</td>
<td>18.5%</td>
<td>90.8%</td>
</tr>
</tbody>
</table>

#### 1.3 Ranking drug-CV pairs from CVAROD

Drug-CV pairs extracted from CVAROD may contain false signals. We compared six ranking algorithms in prioritizing extracted drug-CV pairs using drug-CV pairs extracted from the FDA drug labels as goldstandard. In our previous study in extracting drug-CV pairs from FAERS, we demonstrated that ranking by frequency significantly improved the precisions of top-ranked pairs as compared to other statistical ranking methods including RRR, ROR, RR, IC. Here we investigate whether the same is true in ranking drug-CV pairs extracted from CVAROD. As shown in Fig.2, ranking based on frequency is more effective than ranking based upon the other five approaches. The precision of top-ranked pairs (at recall = 0.1) is 0.16 as measured with drug-CV pairs extracted from FDA drug labels as gold standard. This still low precision indicates that it is likely that many true positives among the top-ranked pairs from CVAROD have not yet been captured in FDA drug labeling. This is reflected by the modest overlap between pairs extracted from FDA drug labels and pairs extracted form CVAROD (Table 3).

#### 1.4 Manual curation of drug-CV pairs extracted from MEDLINE sentences

We manually curated all 1,080 drug-CV pairs extracted from MEDLINE sentences. We did not curate the 1,389 pairs that only co-occurred in MEDLINE abstracts but not in sentences. Among the 1,080 pairs, 525 pairs (48.6%) were true positives (“drug CAUSE CV”), 356 pairs (32.9%) were “drug TREAT CV” pairs, and 199 pairs (18.5%) had no obvious semantic relationships (“drug NONE CV”) (Table 4). Moreover, among the 525 true positives, 477 pairs (90.8%) were not included in FDA drug labels.

#### 1.5 Positive relationship between shared CV events and shared drug target genes

We investigated whether drug-drug pairs that shared CV events also tended to share gene targets. A positive relationship between these two entities could indicate a causal relationship between drug targets and the observed cardiovas-
cicular adverse events associated with targeted anticancer drugs. This indication could in turn open up the possibility of predicting unknown CV events by systematically studying their gene targets and discovering novel drug off-targets, such as targets related to cardiovascular systems, instead of tumor cell growth, based on observed cardiovascular events. As shown in Figure 3, there exists a positive relationship between shared CV events and shared gene targets for drug-drug pairs. For instance, the average number of shared gene targets was 1.678 for all drug-drug pairs (cut off \( \geq 0 \)). The number significantly increased to 2.122 for drug-drug pairs that shared at least 250 CVs (cutoff \( \geq 250 \)) and to 3.857 for drug-drug pairs that shared at least 450 CVs (cutoff \( \geq 450 \)).

Since drug metabolism is responsible for many known drug-related adverse events, we then investigated whether the observed CV events were related to drug metabolism. Any positive correlations between CV events and drug metabolism genes could open up the possibility of predicting drug-associated CV events in specific cancer patients based on their metabolism genotypes (personalized cancer care). As shown in Figure 4, there is a strong positive correlation between CV events and drug metabolism genes. Drug-drug pairs that shared more CV events tended to also share more metabolism genes. The average number of shared metabolism genes for all drug-drug pairs (cutoff \( \geq 0 \)) was 0.39. The number significantly increased to 0.583 for drug-drug pairs sharing at least 250 CV events (cutoff \( \geq 250 \)) and 1.375 for pairs sharing at least 450 CV events (cutoff \( \geq 450 \)). In summary, both gene targets and metabolism genes positively correlated with targeted cancer drug-associated CV events, indicating that the observed CV events may have discoverable genetic causes and that we can predict unknown cardiotoxicities and achieve personalized cancer care by systematically studying drug-associated gene targets and metabolism.

2 Discussion

In this study, we built a comprehensive CV toxicity knowledge base for targeted cancer drugs (tcTKB) by extracting drug-CV pairs from five large scale and complementary data sources. We manually curated all drug-CV pairs that appeared in MEDLINE sentences. We systematically analyzed the correlations between the observed CV events and drug-associated gene targets and demonstrated that tcTKB, in combination with other drug-related data, represents a unique knowledge base for our understanding targeted anticancer drugs and their associated CV adverse events.

Nonetheless, our study has several limitations and can be further improved in the future. First, even though tcTKB includes many drug-CV pairs, it remains unknown what the actual precision and recall of these pairs are. For example, it is difficult to measure how many of the drug-CV pairs in tcTKB are true positives (including both known and unknown positives) and how many of them are true negatives. In addition, even though we used four large datasets in constructing tcTKB, we don’t know what the coverage of this knowledge base is. There may be drug-CV pairs in existence in other data sources, such as patient electronic health records (EHRs).
Figure 4: Relationship between shared CV events and shared metabolism genes.

Second, we observed positive relationship between CV events and drug-associated gene targets. As was previously discussed, the drug-CV pairs in tcTKB may contain noise, which can affect the observed correlations. However, we expect that the noise shall not systematically correlate with drug gene targets. Systems approach to studying drug-CV associations in tcTKB, as done for the widely used protein-protein interaction (PPI) data (which also contains noise and is largely incomplete) may generate many insightful biological hypotheses.

Third, in order to further stratify drug-CV pairs in tcTKB by patient characteristics, we need detailed information about patient characteristics. Patient demographics information can be extracted from MEDLINE abstracts [22], full-text JCO articles as well as FAERS. In the future, we will enrich tcTKB with patient demographics information.

Last but not least, we only extracted cardiovascular events associated with targeted cancer drugs in this study. In the future, we will extract richer sets of toxicities that are associated with targeted cancer drugs, including neurotoxicities, nephrotoxicities, heptotoxicities, and hematotoxicities, and develop systems approaches to studying these observed drug phenotypes in order to understand the molecular mechanisms underlying these toxicities.

Acknowledgement

Xu and Wang have jointly conceived the idea, designed and implemented the algorithms and prepared the manuscript. We would like to thank the three curators from ThinTek for the manual curation.

Funding

RX was supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under the NIH Director’s New Innovator Award number DP2HD084068, the Training grant in Computational Genomic Epidemiology of Cancer (CoGEC) (R25 CA094186-06), and the Grant #IRG-91-022-18 to the Case Comprehensive Cancer Center from the American Cancer Society.

References

Barriers and Facilitators to Patient-Provider Communication When Discussing Breast Cancer Risk to Aid in the Development of Decision Support Tools

Haeseung Yi, MPA1; Tong Xiao, MPH1; Parijatham S. Thomas, MD2; Alejandra N. Aguirre, MPH1; Cindy Smalletz, MPH1; Jill Dimond, PhD3; Joseph Finkelstein, MD, PhD1; Katherine Infante, BA1; Meghna Trivedi, MD1; Raven David, BA1; Jennifer Vargas, BA1; Katherine D. Crew, MD MS1; Rita Kukafka, DrPH MA, FACMI1

1Columbia University, New York, NY; 2MD Anderson Cancer Center, Houston, TX; 3Sassafras Tech Collective, Ann Arbor, MI

Abstract

The purpose of this study was to identify barriers and facilitators to patient-provider communication when discussing breast cancer risk to aid in the development of decision support tools. Four patient focus groups (N=34) and eight provider focus groups (N=10) took place in Northern Manhattan. A qualitative analysis was conducted using Atlas.ti software. The coding yielded 62.3%-94.5% agreement. The results showed that 1) barriers are time constraints, lack of knowledge, low health literacy, and language barriers, and 2) facilitators are information needs, desire for personalization, and autonomy when communicating risk in patient-provider encounters. These results will inform the development of a patient-centered decision aid (RealRisks) and a provider-facing breast cancer risk navigation (BNAV) tool, which are designed to facilitate patient-provider risk communication and shared decision-making about breast cancer prevention strategies, such as chemoprevention.

Introduction

Breast cancer is the most common cancer among women in the U.S. Known risk factors include family history, BRCA genetic mutations, benign breast disease, and reproductive history1. The Gail model or breast cancer risk assessment tool (BCRAT) may be used to calculate a woman’s absolute 5-year and lifetime risk of breast cancer, based upon her age, race, reproductive history, family history, and benign breast disease2. A woman is considered high risk of developing breast cancer if her 5-year risk is ≥ 1.67% or lifetime risk is ≥ 20%. Chemoprevention refers to taking medications for the primary prevention of cancer. Anti-estrogens, such as selective estrogen receptor modulators (SERMs) and aromatase inhibitors (AIs), have been shown in randomized controlled trials to reduce breast cancer incidence by up to 50-65% among high-risk women3,7. Based upon this evidence, the U.S. Preventive Services Task Force and other professional organizations recommend that clinicians discuss chemoprevention with high-risk women4-10. An estimated 15% of women, age 35-70 years, in the U.S. may be eligible for chemoprevention11, but fewer than 5% of high-risk women offered an anti-estrogen agree to take it12. Reasons for low chemoprevention uptake include lack of routine breast cancer risk assessment to identify high-risk women, insufficient knowledge about anti-estrogens on the part of clinicians and patients, and multiple competing demands in the primary care setting12,13. Discussions about chemoprevention represent an opportunity for risk communication and shared decision-making (SDM) to elicit patient preferences and enhance patient-provider communication. Because there is limited knowledge on the part of patients and providers about breast cancer risk assessment and chemoprevention, we developed decision support tools targeting both groups, which will be integrated into clinic workflow. The tools we are developing are a patient-centered decision aid (DA), RealRisks, and a provider-facing breast cancer risk navigation (BNAV) tool to increase breast cancer chemoprevention in the primary care setting.

A study that conducted a systematic review on patient-reported barriers and facilitators to SDM concludes that it is not that patients do not want to play a role in SDM, but it is because they can’t due to various structural, predisposing, interactional, and preparatory factors. The authors suggest that patient-reported barriers should be considered with provider-reported barriers for intervention and implementation14. The purpose of this study was to identify barriers and facilitators in patient-provider communication when discussing breast cancer risk. In order to obtain new information and learn from patient/provider perspectives, we conducted separate focus groups with patients and providers to collect qualitative data. By integrating the results we obtain from this study to the decision support tools, we will be able to further develop the tools to facilitate communication in patient-provider encounters and possibly allow SDM to take place.
Methods

Participants

In June 2013, we conducted four patient focus groups among English-speaking women recruited from Northern Manhattan in New York, NY. Women who participated in a community database through the Columbia Community Partnership for Health were contacted via email or telephone. Each focus group consisted of 7-9 women. A total of 34 women that reside in the Washington Heights/Inwood community participated. Eight key informant interviews of primary care providers (PCPs) of 1-3 participants each (N=10) were conducted at Columbia University Medical Center (CUMC). At CUMC, PCPs practice at 6 locations of the Ambulatory Care Network (ACN) clinics in Northern Manhattan. The patient population is over 80% Hispanic or African American with a predominantly Medicaid/Medicare payer mix. All participants provided written informed consent to audio-recording of a 90-120 minute facilitated session and completion of a brief survey. Patients and providers were given modest incentives of $40-$50 for participation. The study was approved by the institutional review board at CUMC.

Description of the RealRisks Decision Aid

RealRisks is a patient-centered DA that models patient-provider dialogue and incorporates experience-based dynamic interfaces to communicate numeric and probabilistic concepts that are central to breast cancer risk and chemoprevention. It is designed to improve: 1) accuracy of risk perceptions; 2) self-efficacy to engage in a collaborative dialogue about breast cancer risk and chemoprevention; and 3) decision satisfaction. The narrative is based on a fictitious character named Rose, who engages in discussions about breast cancer risk with family, friends, her PCP, and specialists. We have segmented the narrative into the following modules: 1) Risk (what is risk, breast cancer risk factors); 2) Chemoprevention (chemopreventive agents, risks/benefits). RealRisks will be tailored to a woman’s risk, so she will be reviewing only the modules that are most relevant to her. Embedded within the narrative of RealRisks are games of experience-based risk interfaces, based upon our previous work. For example, the first game is about breast cancer risk for an average 50-year-old woman and conveys how time (5-year, lifetime) affects risk with a pictograph of 100 clickable women. Players are instructed to click until they ‘find’ a woman with breast cancer. Players continue to click (e.g., sample from the population of 100 women) to better learn the meaning of a given pre-set probability (e.g., 12 out of 100 women or 12%). A similar game will be adopted in the chemoprevention module to accurately represent the benefits of chemoprevention and the risks of side effects. Accuracy of risk perceptions is important to informed decision-making given that patients may over-estimate their breast cancer risk or the risks of side effects to chemopreventive agents, such as tamoxifen.

Description of the BNAV tool

BNAV uses a two-pronged approach to improve knowledge among PCPs on chemoprevention. One component is the web-based chemoprevention toolbox, a repository of information and resources that is modeled based on the Theory of Planned Behavior. It includes: 1) standard guidelines and a self-paced interactive educational guide (attitudes); 2) video testimonials and a social component that includes the ability to compare their performance against aggregate, anonymous data of their peers (subjective norm); 3) a repository of their patients’ breast cancer risk assessments, along with the action plans based upon their patients’ interactions with RealRisks (perceived behavioral control). Based on appointment scheduling data, a provider would receive a periodic notice of upcoming patients that meet high-risk criteria so as to encourage access to the chemoprevention toolbox, which will be sent as an email or text. The second component of BNAV is embedded within the electronic health record (EHR). Within the ambulatory care dashboard in the EHR used by our PCPs, flagged alerts of patients that meet high-risk criteria for breast cancer with their personalized risk profiles will appear with their mammogram results.

Conducting the Focus Groups

Skilled facilitators (ANA, KDC, RK) led the focus groups using detailed guides. For the patient focus groups, the discussion guide included questions on breast cancer risk factors, BRCA genetic testing, chemoprevention, and discussing breast cancer risk or genetic testing with providers. The discussion guide for the provider focus groups covered questions about genetic testing, chemoprevention, EHR, and communicating breast cancer risk with patients. All sessions were audi-taped.

Data Analysis

For the qualitative analysis, two investigators (HY and TX) independently read the transcript from the first completed patient and provider focus groups to develop the initial codes and coding templates. We identified meaningful segments within the responses and assigned codes using an editing style analysis. Discrepancies in
coding were negotiated at weekly research meetings. HY and TX independently read and coded the remaining focus group transcripts, applying the coding template, which was iteratively modified as the analysis proceeded. We grouped codes into general themes and discussed the themes among the entire team of investigators. The team collectively selected the themes and representative quotes we presented in this paper. Atlas.ti 7.0 software (Atlas.ti GmbH, Berlin, Germany) was used to facilitate qualitative data management and analysis. All transcripts were uploaded into the software to enable investigators to do coding, build the codebook, and group the codes into themes. A final comparison of coding across patient interviews yielded 62.3%-94.5% agreement.

Results

Participant Characteristics

Patients

The majority (61.8%) were Hispanic and mean age was 53.4 years (range, 35-75). Forty-one percent met criteria for low numeracy, defined as a score of 0-5 (range, 0-9)\textsuperscript{30}. Everyone had access to the internet, including 88% who regularly used a computer. In terms of breast cancer risk factors, 8 (23.5%) women had a first-degree family history of breast cancer and 4 (12.9%) had a prior benign breast biopsy. According to the BCRAT (excluding 3 women with a history of breast cancer), mean absolute 5-year and lifetime risk were 1.11% (range, 0.2%-4.3%) and 7.46% (range, 2.8%-14.6%), respectively, and 3 women (9.7%) met high-risk criteria for breast cancer (≥1.67% 5-year risk).

Table 1. Baseline characteristics of patients.

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Patients Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age, years (range)</td>
<td>53.4 (35-75)</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>2 (5.9%)</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>8 (23.5%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>21 (61.8%)</td>
</tr>
<tr>
<td>Asian</td>
<td>1 (2.9%)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (5.9%)</td>
</tr>
<tr>
<td>Numeracy</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>20 (58.8%)</td>
</tr>
<tr>
<td>Low</td>
<td>14 (41.2%)</td>
</tr>
<tr>
<td>Regularly uses computer</td>
<td>30 (88%)</td>
</tr>
<tr>
<td>First-degree family history of breast cancer</td>
<td>8 (23.5%)</td>
</tr>
<tr>
<td>Benign breast biopsy</td>
<td>4 (12.9%)</td>
</tr>
<tr>
<td>High-risk for breast cancer</td>
<td>3 (9.7%)</td>
</tr>
<tr>
<td>Mean 5-year breast cancer risk (SD)</td>
<td>1.11% (0.77)</td>
</tr>
<tr>
<td>Mean lifetime breast cancer risk (SD)</td>
<td>7.46% (2.87)</td>
</tr>
</tbody>
</table>

Providers

To inform the development of BNAV, we conducted individual interviews of 10 physicians affiliated with New York Presbyterian Hospital-Columbia University Medical Center's Ambulatory Care Network (ACN). The majority were female (70%) and they were diverse by race/ethnicity (5 white, 3 black, 1 Hispanic, 1 Asian). They represented various subspecialties in primary care (6 Internal Medicine, 2, Family Medicine, 2 Gynecology) and a range of years in clinical practice (1-35) and years using an EHR (2-10).

Table 2. Baseline characteristics of providers.

<table>
<thead>
<tr>
<th>Provider Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median age, years (range)</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
</tr>
<tr>
<td>African American</td>
</tr>
<tr>
<td>Asian</td>
</tr>
<tr>
<td>Hispanic</td>
</tr>
<tr>
<td>Medical Specialty</td>
</tr>
<tr>
<td>Internal Medicine</td>
</tr>
<tr>
<td>Family Medicine</td>
</tr>
<tr>
<td>Median number of years since completing training (range)</td>
</tr>
<tr>
<td>Median number of years using EHR (range)</td>
</tr>
</tbody>
</table>

**Barriers to communicating risk**

Based on the focus groups, time constraints, lack of knowledge, low health literacy, and language barriers emerged as the main barriers to communicating risk. The results showed that time was considered as the most significant barrier for both patient and provider groups when communicating risk. As shown in Table 3, patients thought providers “don’t have so much time to explain” and “the office visit is very hurried.” They also mentioned not being able to discuss personal circumstances with their providers because of the short amount of time. Providers also talked about the limited time they have with patients. Patients having questions towards the end of the visits was another issue related to time pressure. To worsen the situation, they said that “this encounter with the patient and the physician are getting so long” due to various interest groups trying to take part, which means there is less time to have discussions that the patients and providers would like to have. Patient’s lack of knowledge was discussed as a barrier to both groups. Patients said, “Because we don’t have the knowledge and don’t ask, the doctor won’t give it to you.” Providers seemed to think that patients do not know much about medical terms and risk-related numbers, which makes it difficult to communicate risk. Providers mentioned low health literacy as an obstacle for patient-provider communication. Since providers are not aware of the patient’s health literacy level, they said, “My struggle is how to get this information across without creating more anxiety with the health literacy issues.” The language barrier in a largely Spanish-speaking community in Northern Manhattan was also discussed as an obstacle. Providers experienced difficulty when explaining risk to patients mainly due to language and cultural barriers.

**Table 3.** Quotations on barriers to communicating risk.

<table>
<thead>
<tr>
<th>Barriers</th>
<th>Sample quotations</th>
</tr>
</thead>
</table>
| Time constraints   | *Patient* \  
                      Usually when you go to see the doctor, they don’t have so much time to explain. \  
                      Since the office visit is very hurried, five or seven minutes, he doesn’t have time to go over my personal circumstances. \  
                      Even if you have so many questions, the doctor will come out and tell you that they have other patients to see and they don’t have time. \  
                      *Provider* \  
                      The time in the doctor’s room is pressured. \  
                      A lot of time they don’t realize what it’s about and they bring it up at the end. In that case, when I’ve already spent 45 min with them and we only have 15 min., then it does open this Pandora's box. |
Facilitators to communicating risk

Information needs, desire for personalization, and autonomy emerged as facilitators to patient-provider communication through the dialogue (Table 4). Patients showed strong interest in obtaining more information. They said that this information does not have to be from a doctor and they are willing to learn from various sources (“I think it’s important that you have more places to ask questions about breast cancer, even if that person might not be my doctor. It might be a site, a nurse”). Not only will this information help them understand the doctor’s explanation, but also will help them “feel more comfortable”. They also mentioned having many questions they want to ask when
they are at their doctor’s office (“Half the time when we are at the doctors, we want to ask ten thousand questions”). Providers also had information needs. They said that they are willing to participate in courses that provide information on breast cancer. Desire for personalization was addressed in the patient focus groups. Patients said, “I wouldn’t blindly follow a doctor’s recommendations if it’s a general recommendation. I want it to be individualized.” However, providers point out that patients who do not need certain services (e.g., cancer screening, genetic testing) are receiving them, and those who need them are not. Patients are eager to participate in clinical decision-making and are willing to reach out to other medical groups if necessary. Providers also notice this strong interest in autonomy among patients. They think “this patient population really wants to learn and be involved” during the patient-provider encounters. However, providers seem to be frustrated when patients value doctor’s opinion less and think they know what should be done.

Table 4. Quotations on facilitators and information needs to communicating risk.

<table>
<thead>
<tr>
<th>Facilitators</th>
<th>Sample quotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desire for personalization</td>
<td><strong>Patient</strong>&lt;br&gt;I wouldn’t blindly follow a doctor’s recommendations if it’s a general recommendation. I want it to be individualized. [sic]&lt;br&gt;The doctors are reading off of a script. They recommend the same thing for all of us. […] It doesn’t seem personal at all. <strong>Provider</strong>&lt;br&gt;The other point is again what you’re discussing is a real public health issue on both sides. Women who do not need mammograms are getting mammograms and women who need genetic testing are not getting that. This is a process of education and it’s a long process in that area. The investment has to be done in the high school, PTA, and make mothers and students alike aware that it’s very important.</td>
</tr>
<tr>
<td>Autonomy</td>
<td><strong>Patient</strong>&lt;br&gt;Shouldn’t they do it for whoever wants to have it done?&lt;br&gt;But unfortunately we all need to be very proactive and never accept no for an answer.&lt;br&gt;I would consult not just my doctor but I would also consult some other medical type groups to see what I could do.&lt;br&gt;We all have to remain clear on what we want. How is it that we learn what we want? If you don’t know what breast cancer is, then how do you go to a doctor and say I’d like to get checked for breast cancer. <strong>Provider</strong>&lt;br&gt;I think this patient population really wants to learn and be involved. I’ve worked with other patient populations where they want my opinion less and they know or they think they know the right answer and that was frustrating. I really think this patient population wants to have these discussions.</td>
</tr>
<tr>
<td>Information Needs</td>
<td><strong>Sample quotations</strong></td>
</tr>
</tbody>
</table>

**Patient**<br>I think it’s important that you have more places to ask questions about breast cancer, even if that person might not be my doctor. It might be a site, a nurse.<br>I am more informed so I can have a better discussion and I can understand his responses better. I know more terms. I don’t know if it will get me to talk
about it with my doctor but I will feel more comfortable.
You know, you don’t want to walk away and not knowing.
Half the time when we are at the doctors, we want to ask ten thousand questions, we forget what to ask.

Provider

Facilitator: Yeah so it would be like those 20 minute talks but more relevant to the primary care providers on how to assess risk and who should get genetic testing and that type of thing. So if we made this a requirement to take this course online, do you think that would be helpful? Absolutely. Yeah and I would definitely be interested.

Discussion

In summary, patients and providers consider time, lack of knowledge, low health literacy, and language barriers as obstacles to communicating risk. On the other hand, information needs, desire for personalization, and autonomy are discussed as facilitators that may enhance communication in patient-provider encounters. However, discordance between patient and provider expectations about good clinical practice still exists and may hamper SDM. We will use this information to inform the development of RealRisks and BNAV to facilitate communication about breast cancer risk.

By integrating an education component in RealRisks, the health literacy and lack of knowledge issues can be addressed. Educational materials can be presented in various versions to target patients with different levels of understanding. Including a Spanish version of the material will help overcome language barriers since there are many Hispanic patients. Also, patients can learn about breast cancer through RealRisks whenever they want to, which will not only fulfill their information needs, but also their needs to gain information from sources other than doctor visits. When further developing BNAV, a health literacy indicator could help providers determine the health literacy level of each of the patients, and therefore, result in a level-appropriate approach when explaining risk. In addition, an education toolbox can be implemented to educate providers about breast cancer risk that will inform discussions with their patients.

RealRisks and BNAV can address the time issue that was the biggest concern during both patient and provider focus groups. By interacting with RealRisks, patients can receive education before meeting with providers, which can reduce time spent on providers explaining risk and related concepts. This could allow time for discussing patients’ preferences, which is a component in SDM. Patients’ strong interest in autonomy and personalization will also enhance risk communication and SDM. Integrating patient’s needs and preferences with provider’s expertise and other resources, SDM could help patients make informed medical decisions.

This study has several limitations. The sample size of patients and providers was relatively small. In addition, a large proportion of our study patients were Hispanic, an ethnic group which is often under-represented in clinical studies. Our results may not be generalizable to populations from other geographic regions; however, Hispanics are the largest minority group in the U.S. We also only included physicians and did not collect data from other primary care providers (e.g., physician’s assistants, nurse practitioners), who may be targeted in future studies.

We believe that RealRisks and BNAV can facilitate patient-provider communication and possibly lead to SDM when the barriers are dealt with and facilitators are integrated as discussed. For future studies, since a majority of the patients are Hispanic, it would be interesting to integrate communication facilitators that target this ethnic group. In a study that compared Hispanic and Non-Hispanic women’s needs in patient-provider communication, the results showed that Hispanic women experienced difficulty during the communication process because of language barriers and health literacy, which is consistent with our results. Interestingly, while a warm communication style was important for Hispanic women, information needs were crucial for Non-Hispanic women. It would be meaningful to conduct research on how to include ethnic-specific communication styles into our tools and examine whether the further developed tool was effective in enhancing patient-provider communication.

This research will allow patients to gain a better understanding of their breast cancer risk, along with their PCPs, which will inform risk-based screening and prevention strategies (e.g., genetic testing, chemoprevention). Our goal
is to maximize benefits, minimize harms, and promote more efficient allocation of health services, particularly for high-risk individuals.

References
Evaluating Term Coverage of Herbal and Dietary Supplements in Electronic Health Records

Rui Zhang, PhD$^{1,2}$, Nivedha Manohar, MHI$^1$, Elliot Arsoniadis, MD$^{1,2}$, Yan Wang, MS$^1$, Terrence J. Adam, RPh, MD, PhD$^{1,3}$, Serguei V. Pakhomov, PhD$^{1,3}$, Genevieve B. Melton, MD, PhD$^{1,2}$

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

Abstract
Herbal and dietary supplement consumption has rapidly expanded in recent years. Due to pharmacological and metabolic characteristics of some supplements, they can interact with prescription medications, potentially leading to clinically important and potentially preventable adverse reactions. Electronic health record (EHR) system provides a valuable source from which drug-supplement interactions can be mined and assessed for their clinical effects. A fundamental prerequisite is a functional understanding of supplement documentation in EHR and associated supplement coverage in major online databases. To address this, clinical notes and corresponding medication lists from an integrated healthcare system were extracted and compared with online databases. Overall, about 40% of listed medications are supplements, most of which are included in medication lists as nutritional or miscellaneous products. Gaps were found between supplement and standard medication terminologies, creating documentation difficulties in fully achieving robust supplement documentation in EHR systems. In addition, in the clinical notes we identified supplements which were not mentioned in the medication lists.

Introduction
The popularity of herbal and nutritional supplements in the United States (U.S.) has expanded rapidly in recent years. The American Botanical Council reported that nearly 6 billion dollars are spent on herbal supplements in the U.S. in 2013. The Center for Disease Control reports that around half of U.S. adults use some form of dietary supplements. Recently, the National Center for Complementary and Integrative Health (NCCIH, formerly the National Center for Complementary and Alternative Medicine) published the National Health Statistics Reports reporting that about 40% of Americans use some form of complementary and alternative medicine (CAM) and that nonvitamin/nonmineral dietary supplements were still the most commonly used complementary health approach among adults in 2002 (18.9%), 2007 (17.7%) and 2012 (17.7%). Among these, fish oil supplements (7.8%), glucosamine chondroitin, or a combination (2.6%) were found to be the most commonly consumed supplements (2012 National Health Interview Survey). The use of melatonin and probiotics has also dramatically increased from 0.6% and 0.4% in 2007 to 1.3% and 1.6% in 2012, respectively.

CAM therapies are typically used to complement conventional medicine with the goal of reaching a better healthcare outcome, although less than 5% of U.S. adults exclusively used CAM therapy approaches. One in four persons taking a prescription medicine also take an herbal supplement, increasing the possibility of drug-supplement interactions (DSIs). For example, warfarin can interact with various supplements such as Panax ginseng and Ginkgo biloba leading to severe adverse effects such as spontaneous postoperative bleeding. St. John’s Wort has also attracted attention as it lowers blood concentrations of cyclosporine, amitriptyline, digoxin, indinavir, and warfarin, potentially resulting in severe clinical syndromes. Another clinical study reported that volunteers receiving digoxin and St. John’s Wort had a 25% decrease in the concentration of digoxin in blood plasma. Unfortunately, the ability to readily identify adverse effects of herb and dietary supplements and their reactions with conventional Western medications is not well studied and the reports on such interactions occur less frequently in clinical practice.

DSIs are most often due to an inhibitory or synergistic interactions between drugs and supplements with similar genes or molecular pathways within the body. Before market approval, new drugs are usually tested for interactions with other drugs through pharmacology experiments and clinical studies. Testing for interactions between new drugs and supplements, however, are not required due to the differences in regulatory requirements for supplements, as delineated in the Dietary Supplement Health and Education Act of 1994 (DSHEA). The DSHEA rules designate supplements as food, which requires appropriate labeling and adherence to food safety rules. However, since supplements are not regulated as typical medications, they do not need to meet the usual clinical approval trials for
safety and effectiveness, which includes assessment of medication pharmacology and potential drug-drug interaction characteristics. As such, post-market surveillance through pharmacoepidemiology studies and other standard mechanisms which are often used to detect adverse events in a given population are not required. These methods are limited, moreover, as they can only focus on a small set of drugs or supplements.

Electronic Health Record (EHR) systems are the main communication and documentation platform for healthcare providers. They store a large amount of data on drug prescribing behaviors, adverse drug events, and patient symptoms, thus providing a rich source of observational data offering the potential for active surveillance. Most current efforts use structured data such as diagnoses to find drug interactions and adverse drug events. These experiments often miss important information in clinical notes that can be leveraged for further clinical research and knowledge discovery. Unstructured texts can be analyzed using natural language processing (NLP) techniques for patient cohort identification, phenotype extraction, and drug-related information extraction.

To facilitate effective mining of EHRs for DSIs, a better understanding of how the EHR represents the supplements and how effective the term coverage is for both structured clinical data and unstructured clinical notes is needed to extract accurate and comprehensive supplement information from the EHR. To the best of our knowledge, the investigation of supplement term representation and coverage on EHR systems is limited and deserves further investigation. One possible reason for the lack of extensive studies on DSIs is an absence of a standard and accepted terminology for herbal supplements. In a recent study comparing terms with different resources, we found that none of five major online databases covered all supplement terms. In this study, we sought to evaluate supplement term coverage between online supplement databases and EHR patient data. We also sought to investigate the adequacy of standard terminologies for representing supplements existing in the EHR.

**Background**

In this section, we introduce five online databases used to form a supplement list and the standard terminologies we evaluated for representing supplements.

**Supplement databases**

**Natural Standard Authority Database (NSAD)**

NSAD is a database set up by Natural Standard, a research collaboration including physicians and researchers. There is a grade given to each entry, which reflects the level of evidence-based literature available about each product and its use. Each entry has an overview of the supplements, common names all over the world, uses, warnings, contraindications, mechanism of action, and a literature review. In addition to having entries for each of the supplements, there is an adverse reaction checker, an effectiveness checker, interactions display and a search for pregnancy related information.

**Medscape**

Medscape is managed by the WebMD health professional network; the material is provided by physicians and authorities in that field. Medscape provides author attribution and sources of information making it a reliable source. Medscape also features a list of searchable entries and a tool which allows users to check for drug or supplement interactions.

**Natural Medicines Comprehensive Database (NMCD)**

NMCD is managed by the therapeutic research center. The database has over 1,000 terms, but allows users to search in several other languages. Under each product, the effectiveness of the supplement, safety, known interactions with drugs, mechanism of action, and adverse reactions are listed. The information is targeted at physicians, researchers, and pharmacists. There is a separate advanced search for physicians which can used to make clinical decisions. This database also has an interactions checker and a search tool which lets users search with common names, scientific names or brand names making it one of the most useful and comprehensive databases.

**MedlinePlus**

MedlinePlus is a web-based information service provided by the U.S. National Library of Medicine (NLM). It has a health topic section, a drug and supplements section and several other tools. This service including the drug and supplement database is maintained by the National Institutes of Health and is aimed at the general public. The information is provided by the National Center for Complementary and Alternative Medicine, National Toxicology...
Program, and the Office of Dietary Supplements. Each entry is followed by a journal style list of references for that supplement.

*Drugs.com*22

Drugs.com provides popular supplements with efficacy, side effects and interactions with drugs. The information is targeted at consumers and the format is user-friendly. Drugs.com also has a pill identifier and an interactions checker that lets users search for every potential interaction a drug will have with other drugs, supplements or food. Interactions are further classified as severe, moderate or mild. Information is derived from Wolters Kluwer Health Inc.

**Unified Medical Language system (UMLS) and MetaMap**

UMLS is a repository that integrates over 100 medical vocabularies from many sources and provides a unified platform which can be used to develop or enhance applications. One of the components in the UMLS is a Metathesaurus which has over 2 million terms and codes from many different vocabularies like Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), Current Procedural Terminology (CPT), Logical Observation Identifiers Names and Codes (LOINC), etc. For this paper, UMLS was used to normalize data from different sources (i.e., EHR and online databases). Each concept in the UMLS Metathesaurus has a Concept Unique Identifier (CUI), which can be used to compare data from different sources.

MetaMap is a natural language processing tool developed by the NLM to automatically map biomedical texts to UMLS Metathesaurus. It is able to lexically and syntactically analyze the texts and provide a list of mapping concept candidates with mapping scores.

**RxNorm, NDF-RT, and RxMix**

The NLM creates a standardized nomenclature of clinical drugs called RxNorm from a source of 12 drug vocabularies. RxNorm has unique identifiers for each drug along with the drug’s dosage, generic name, chemical components, and dosage forms.23

National Drug File - Reference Terminology (NDF-RT) is a part of the Veteran Health Administration’s National Drug File. It is able to classify drugs into formal categories in addition to giving information about their molecular interactions, kinetics, therapeutic categories, and dose forms.24

RxMix is a web application that allows users to combine various functions from the RxNorm and NDF-RT application program interfaces (APIs) to create custom applications that can be run interactively or in a batch mode.25 We used this tool for mapping terms to RxNorm and NDF-RT concepts.

**Methods**

We first extracted supplement terms from both online databases and structured medication lists in EHRs, and then formed a comprehensive list of supplements followed by a search for the terms in clinical texts. All terms from online resources, medication lists, and clinical texts were mapped to UMLS, RxNorm and NDF-RT. Then, we compared the term coverage for the three resources (Figure 1).

**Extraction of supplement list from online databases**

To identify the selected databases, we performed an Internet search for the top herbal supplements databases to identify the selected databases. Any databases without support by evidence or with very short (<50) supplement lists were excluded. Databases no longer updated or maintained were also rejected for this study. We finally selected the five popular online databases: NSDA, Medscape, NMCD, MedlinePlus, and Drugs.com as the sources of herbal and dietary supplement terms. We extracted all supplement terms from databases and manually excluded the non-English words from online databases. A combined list of supplements was finally generated from online databases for further analysis.

**Extraction of supplement terms from clinical data**

We collected patients’ medication lists from the University of Minnesota Medical Center, Fairview Health Services (FHS) during a four-year period of 2011-2014. University of Minnesota institutional review board approval was obtained and informed consent was waived for this minimal risk study.

The collected medication list from FHS were processed by using the following steps:
• Step 1: A physician (coauthor EA) manually reviewed medication lists to select supplement terms;
• Step 2: Supplements related information such as the frequency of supplement usage and their assigned pharmaceutical class (e.g., Nutritional Products), were extracted and analyzed to provide a better understanding of how supplements are represented in the medication lists;
• Step 3: The list of supplements was mapped to the UMLS by using MetaMap. Only exact matches with a perfect score of 1000 were considered and retrieved;
• Step 4: The supplement list was also mapped to standard terminologies RxNorm and NDF-RT by using RxMix25. We specifically used the RxNorm:findRxcuiByString and NDF-RT:findConceptsByName functions to extract the concepts.

Collection of supplement terms from clinical texts
We collected clinical texts from FHS during the same period, and analyzed the clinical texts using the following steps:
• Step 1: Instead of manually reviewing charts to find the supplements, we first combined the supplement lists extracted from both online databases and medication lists to form a comprehensive list;
• Step 2: The comprehensive list (blue dashed lines in Figure 1) was then used as a dictionary to search all clinical texts to investigate if these terms existed in the clinical texts;
• Step 3: To normalize the supplement terms, we mapped these supplement terms to UMLS, RxNorm and NDF-RT, the same as steps 3&4 when mapping medication list in the above section.

Evaluation of supplements term coverage
We evaluated the term representation in medication lists based on their frequency and assigned classes. After supplement terms were mapped to UMLS, RxNorm and NDF-RT, we then evaluated the gap between the current
terminologies to the existing terms. We used Venn diagrams to compare the term coverage of these concepts to see how they overlap with each other to provide a visual and quantitative assessment.

Results

Extraction of supplement list from online databases, clinical data and clinical texts

We extracted 3,115 unique supplement terms from the five online databases, 3,720 unique terms from medication lists, and 4,717 terms from clinical notes.

In the whole medication list in the EHR, we found about 40% of the listed medications are related to supplements. Among these supplements extracted from the EHR medication lists, most of them were classified as Nutritional Product (41.7%) and Miscellaneous Products (40.7%). Other types include Gastrointestinal Agents, Hematological Agents, Gastrointestinal Agents, and other types (Table 1). The top frequent supplements in the medication list include Fish oil, Glucosamine chondroitin, Probiotics, Melatonin, Coenzyme Q10, etc.

Table 1. The supplement representation in medication list in EHR; percentages and classes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUTRITIONAL PRODUCTS</td>
<td>41.7%</td>
</tr>
<tr>
<td>MISCELLANEOUS PRODUCTS</td>
<td>40.7%</td>
</tr>
<tr>
<td>GASTROINTESTINAL AGENTS</td>
<td>8.0%</td>
</tr>
<tr>
<td>HEMATOLOGICAL AGENTS</td>
<td>7.7%</td>
</tr>
<tr>
<td>GASTROINTESTINAL AGENTS</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

Evaluation of supplements term coverage

As shown in Figure 2, after mapping these terms to UMLS and specific standard terminologies, we obtained 1,683, 897, and 1,937 UMLS concepts, 556, 1,410, and 1,733 RxNorm concepts, and 421, 367, and 576 NDF-RT concepts from online databases, medication lists and clinical texts.

Figure 2. Number of terms or concepts in each resource after mapping to the UMLS, RxNorm and NDF-RT.
For all three resources, over 60% of terms cannot be mapped to UMLS concepts, and even much less to RxNorm and NDF-RT. The number of mapped UMLS concepts is larger than RxNorm and NDF-RT for online databases and clinical notes, but the EHR medication list returns more RxNorm concepts than UMLS and NDF-RT concepts.

After mapping to the various terminologies and comparing the concepts coverage between three resources, we found that all concepts extracted from the medication list were a subset of those in the clinical notes. The concepts in both the medication list and the clinical notes have different degrees of overlap with those in online databases. In other words, clinical notes contain all supplement concepts in the medication list and part of the online databases while online databases do not contain all concepts existing in the EHR.

Figure 3. Venn diagram showing the supplement term coverage between online databases, medication list, and clinical notes after mapping to (A) UMLS concepts, (B) RxNorm concepts, and (C) NDF-RT concepts.

We further investigated those supplements, in clinical notes but did not include them in the supplement list (shown in Figure 3), and listed a few selected supplements with their occurrences in example sentences (Table 2). Many of them are related to history of food consumption, allergy problems and recommendation to take or avoid a specific supplement either due to risk of potential drug interactions or due to the known physiologic effects of the supplements. But we also found that patient was actually taking or using that supplement, such as Goat’s rue.
Table 2. Supplements only mentioned in clinical texts and not in the medication lists. Supplements are underlined in the examples.

<table>
<thead>
<tr>
<th>Supplements</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aloe vera</td>
<td>We discussed avoiding <em>aloe vera</em> for the potential interaction with cyclosporine.</td>
</tr>
<tr>
<td></td>
<td>… patient notes ‘irritation and sensitivity’ of skin, discussed use of 100% <em>aloe vera</em> gel mixed with Aquaphor, Mepilex given to patient for comfort along left axilla and instructed on use.</td>
</tr>
<tr>
<td></td>
<td>… pt has also started on <em>aloe vera</em> juice.</td>
</tr>
<tr>
<td>Goat’s rue</td>
<td>she is taking her <em>goat’s rue</em> and mother’s milk plus supplements.</td>
</tr>
<tr>
<td></td>
<td>Discussed herbals and mom is taking both Fenugreek and <em>goat's rue</em> to increase alveoli.</td>
</tr>
<tr>
<td></td>
<td>She tried herbals including <em>goat's rue</em> to increase alveoli without success.</td>
</tr>
<tr>
<td>White oak</td>
<td>Allergens positive for cat and dog dander, <em>white oak</em> and ragweed.</td>
</tr>
<tr>
<td></td>
<td>… a 2/6 reaction to American elm (white), and a 1/6 reaction to <em>white oak</em> and ragweed.</td>
</tr>
<tr>
<td></td>
<td>Approximately 3 days ago he scraped his left anterior leg with a piece of white oak.</td>
</tr>
<tr>
<td>Marshmallow</td>
<td>Continue <em>aloe vera</em> and <em>marshmallow</em> root tea supplements.</td>
</tr>
<tr>
<td></td>
<td>SORE THROAT 1) gargle with salt or apple cider vinegar mixed with water  2) tea (<em>throat coat</em> tea includes licorice, <em>marshmallow</em> root and slippery elm or you can make cinnamon and honey tea - cinnamon has analgesic properties).</td>
</tr>
<tr>
<td>Adenosine</td>
<td>We then performed thrombectomy of the distal RCA after using nicardipine, nitroglycerine and <em>adenosine</em>.</td>
</tr>
<tr>
<td></td>
<td>Should and EKG that showed no other acute abnormalities he was given 6 mg of <em>adenosine</em> with resolution of the SVT.</td>
</tr>
<tr>
<td></td>
<td>No significant abnormalities were noted with infusion of <em>adenosine</em>.</td>
</tr>
<tr>
<td>Buckwheat</td>
<td>Honey has anti-inflammatory properties (darker honey such as <em>buckwheat</em> is the best, give it on the spoon or make chamomile tea (which is also anti-inflammatory) with LOTS of honey).</td>
</tr>
<tr>
<td>Hazelnuts</td>
<td>Have spinach (cooked or raw), colorful fruits, walnuts, <em>hazelnuts</em>, almonds in your diet.</td>
</tr>
<tr>
<td></td>
<td>Food intolerance Oral allergy syndrome: with <em>hazelnut</em> and birch pollen--pt eats all other nuts without a problem.</td>
</tr>
<tr>
<td></td>
<td>… 1-2 ounces (a small handful) of almonds, walnuts, <em>hazelnuts</em> or pecans once a day in place of other less healthy snacks.</td>
</tr>
<tr>
<td>Spearmint</td>
<td>Applied seaband, <em>spearmint</em> aromatherapy, Reiki therapy</td>
</tr>
<tr>
<td></td>
<td>Therapies tried and outcome: Baby lotion no relief, Mary Kay lotion with <em>spearmint</em> with moderate relief.</td>
</tr>
<tr>
<td></td>
<td>Try limiting chocolate, peppermint, and <em>spearmint</em>.</td>
</tr>
<tr>
<td>Black tea</td>
<td>Patient took some oral liquids while in clinic (<em>black tea</em>), and tolerated this well.</td>
</tr>
<tr>
<td></td>
<td>Drinking green tea, will switch to <em>black tea</em>.</td>
</tr>
<tr>
<td>Sitostanol</td>
<td>Plant sterols such as Benecol: Three servings per day (1.5 g <em>sitostanol</em> per 1 1/2 teaspoon [8 g serving])…</td>
</tr>
<tr>
<td></td>
<td>Consuming plant sterols such as beta-sitosterol and -<em>sitostanol</em> (typically found in margarine spreads such as Promise activ or Benecol).</td>
</tr>
<tr>
<td>Poppy seeds</td>
<td>His wife works with natural foods and he does eat a lot of seeds including <em>poppy seeds</em>.</td>
</tr>
</tbody>
</table>
It is reasonable that the low level of morphine in his urine could be from high levels of seed intake including poppy seeds and this does not change my plan for opiate analgesics.

| Quinoa  | … organic meat, brown rice or quinoa or oats (gluten free), amaranth, plain steamed veggies, apples, grapes. Try whole grains such as 7-grain breads, whole-wheat pasta, brown rice and other grains such as quinoa, barley, oats, and millet. |

Discussion

DSIs are attracting more attention recently due to their known and potential adverse effects on patient safety. Many clinical studies have found potential interactions, although they focused on a small subpopulation. In addition, use of EHR data for DDI and drug adverse effects has been widely investigated, while only limited study has focused on mining EHR data to explore DSIs. This research investigates how EHR data represent herbal and dietary supplements and evaluates existing terminologies to represent the supplements in EHR systems. This paper studies DSIs by first establishing a list of supplements from a variety of sources and then identifying the coverage of the supplements in the medication lists and notes of an EHR system.

It is not required to document supplement usage in the EHR system, so it is surprising to find that about 40% of the medications documented in the EMR are herbal and dietary supplements. Unsurprisingly, the most frequently used nonmineral/nonvitamin supplements in the medication lists are consistent with the top 10 supplements listed in a recent report of US consumption of supplements.4 In the medication list, there is no easy way to identify the supplements other than manual review. The reason is that there is not a class called herb or dietary supplements or similar, and most of them are classified as nutritional or miscellaneous products. There are still 3% of nutritional products and 40% of miscellaneous products that are non-supplements.

The supplement names are similar to drugs in that they have many ways to represent them with common names or trade names. Many names in the medication lists are the combinations of two or more ingredients, connected by either “-”, “/o” and/or “w/”, such as “Alpha Lipoic Acid-Cr-Cinnamon”, “Zinc Citrate-Phytase”, “Prenat-FeFum-FePo-FA-Omega 3”, “Prenat w/o A-FeCbGI-DSS-FA-DHA”, and “B-Complex w/ C-Biotin-D-Zinc & Folic Acid”. Many names also contain information of enteral formulation such as “Tab” (i.e., Tablet), “Cap” (i.e., Capsule), parenteral formulations, such as “(Bulk) Powder”, “(Bulk) Granules”, “Crystals”, “Liquid”, and dosage information such as “200 MG”, “100 MG/ML”. These supplements with dosage and formulation information and those containing multiple ingredients were neither mapped to the UMLS concepts nor mapped to other terminologies. We did not perform a normalization step before mapping to keep some detailed information which can be recognized by RxNorm. Moreover, separating the combined supplement names into each of the ingredients is not meaningful. The RxNorm was designed to identify drugs with detailed information, so the terms with different dosages can be linked to their unique RxNorm concepts, although missed mappings still exist. For example, two supplements with the same ingredients and different dosages “Alpha Lipoic Acid (Thioctic Acid) Cap 200 MG” and “Alpha-Lipoic Acid (Thioctic Acid) Cap 300 MG” were mapped to the same UMLS concept “C0023791:.ALPHA.-L IPOIC ACID [Organic Chemical,Pharmacologic Substance,Vitamin]”, but they have their own unique RxCUIs: 313841 and 333831, respectively. This is one reason why the number of mapped unique UMLS concepts is less than RxNorm concepts for supplements in the medication lists. Another reason is that many terms cannot find suitable UMLS concepts, but can be related to RxNorm concepts. For example, “Cinnamon Tab 500 MG” does not have an exact match in UMLS but does have a unique RxCUI of 6459.

A gap between terminologies and supplements in EHR was also observed. Terms with multiple ingredients cannot be mapped and were treated differently from those listed with one of these ingredients. Many terms with detailed dosage information can only be partly mapped to UMLS concepts with a mapping score lower than 1000, in which cases we treated them as non-mapped terms. NDF-RT has less coverage than the other two for the supplements. Not only for those terms with multiple ingredients or detailed information but also some general supplement terms such as “Chia Oil” and “Garlic” which cannot be mapped to NDF-RT. Thus, none of these terminologies can be used to represent all the supplement terms in the EHR system.

Considering many forms of a single supplement name, we mapped and normalized them to UMLS, RxNorm, and NDF-RT concepts before comparing the term coverage. By searching supplement lists both in online databases and medication lists, we found that all terms in the medication list were included in the list generated from clinical notes,
which makes sense as clinicians usually import the current medication list as part of the clinical notes. About 64% of terms in the online databases were mentioned in the clinical notes, indicating the wide consumption of supplements. The lower overlap of UMLS concepts between the medication lists and major online databases may be due to the mapping issues mentioned above. But they have higher overlap 37% in RxNorm than 26% in UMLS concepts. The possible reason is that RxNorm is more sensitive to the supplements with detailed information than UMLS.

There are some supplements in clinical notes that are not mentioned in the medication lists. For example, “We discussed avoiding aloe vera for the potential interaction with cyclosporine” indicates the clinician realized the potential DSI between aloe vera and cyclosporine and suggested the patient to not use aloe vera. A mother also tried an herb supplement called goat’s rue to increase alveoli. The supplement marijuana was also mentioned as a clinical history component and relates to the patient’s strange behavior. Adenosine shows up in the notes but not in the medication list as it is a provider administered medication used as an acute therapy. Such medications may only show in the documentation as provider orders or as clinical note documentation as they would not have a role as a chronic medication.

We also found similar terms but not exactly the same names in both clinical notes and medication list. One example, “Cranberry juice” is in the clinical notes, but the medication list only has “CRANBERRY JUICE EXTRACT PO”, which does not exactly match with the UMLS concept “C1572601:CRANBERRY JUICE [Food]” (score of 694). Since we only include supplements with exact matches, we excluded this from the supplement list extracted from the medication list. Although some supplements existed in both sources but notes provide additional information; for example, “[ginkgo] increased risk for bleed especially since taking aspirin - reviewed with pt.” infers potential interactions between ginkgo and aspirin. Other examples include “Ginkgo has been shown to inhibit hepatic glucuronidation of MPA in vitro.” and “Medication Changes: He had been taking ginkgo but read an article that said it should not be taken with warfarin so he stopped taking it about 2 days ago.” This also suggests to us that it is necessary to complement unstructured data with structured data for obtaining comprehensive information from the EHR.

This pilot study has multiple limitations. We only collected clinical data from one site – FHS, during the limited period of 4 years, which may underestimate the term coverage in the EHR. When mapping to UMLS, we only considered the exact matches, so this brings lower returned concepts from UMLS. We did not distinguish reported alimentary facts from medical facts at the current stage. Moreover, it may have bias when we searched supplements in clinical notes by comparing with the supplements generated from the other two sources. The best way is to manually review, but that would be costly in terms of time and labor.

Conclusion

We evaluated supplement term coverage in the EHR by comparing with the existing online resources. Forty percent of medication lists are supplements and they are mostly categorized as Nutritional Products or Miscellaneous Products. All supplements in the medication lists were also mentioned in the clinical notes. We found there is a gap between standard terminologies and supplements in the EHR and not a single standard terminology can cover all supplements in the EHR. Moreover, clinical notes contain additional information such as suggestion or discussion of supplements existing in the medication lists as well as those supplements not mentioned in the medication lists.

Acknowledgments

This research was supported by the University of Minnesota Informatics Institute on the Horizon grant (RZ), the Agency for Healthcare Research & Quality grant (R01HS022085) (GM), and University of Minnesota Clinical and Translational Science Institute supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR000114) (Blazar). The authors thank Fairview Health Services for their support of this research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Agency for Health Research & Quality.

References

Handling Temporality of Clinical Events for Drug Safety Surveillance

Jing Zhao¹, MS, Aron Henriksson¹, MS, Maria Kvist¹,², PhD, MD, Lars Askér¹, PhD, Henrik Boström¹, PhD
¹ Dept. of Computer and Systems Sciences, Stockholm University, Sweden
² Dept. of Learning, Informatics, Management and Ethics, Karolinska Institute, Sweden

Abstract

Using longitudinal data in electronic health records (EHRs) for post-marketing adverse drug event (ADE) detection allows for monitoring patients throughout their medical history. Machine learning methods have been shown to be efficient and effective in screening health records and detecting ADEs. How best to exploit historical data, as encoded by clinical events in EHRs is, however, not very well understood. In this study, three strategies for handling temporality of clinical events are proposed and evaluated using an EHR database from Stockholm, Sweden. The random forest learning algorithm is applied to predict fourteen ADEs using clinical events collected from different lengths of patient history. The results show that, in general, including longer patient history leads to improved predictive performance, and that assigning weights to events according to time distance from the ADE yields the biggest improvement.

Introduction and Motivation

Drug safety is a critical public health issue, with the prevalence of hospital admissions caused by adverse drug events (ADEs) ranging from 2.4% to 12.0%¹,²,³. Although the safety of a drug must be evaluated in clinical trials prior to its launch, there are potentially many undiscovered side effects due to limitations in terms of both the number of participants and follow-up time. In fact, many drugs have been withdrawn from the market for causing fatal ADEs, such as Vioxx for its doubled risk of causing myocardial infarction⁴ and Cerivastatin for causing fatal rhabdomyolysis⁵. The continued surveillance of drug safety, post marketing, is therefore of utmost importance. Currently, this activity is almost exclusively dependent on spontaneous reports⁶, i.e., voluntary reporting of ADEs by patients or clinicians. Research on using alternative data sources, such as electronic health records (EHRs), are, however, emerging⁷,⁸,⁹. The use of EHRs for pharmacovigilance has several advantages that can complement and compensate for some of the weaknesses of spontaneous reports, such as under-reporting of ADEs and limited access to patient history¹⁰. An important property, and advantage, of EHRs is that they record longitudinal observations of patients and their treatment, including drug use. This opens up great opportunities for analyzing the temporal relationships between ADEs and other clinical events, which, in turn, can aid in the detection of ADEs.

In comparison to spontaneous reports, the reporting of ADEs in EHRs is better regulated, yet the problem of under-reporting remains¹¹. To improve the reporting of ADEs, given the large amounts of data archived in EHR systems, manually screening all records to identify ADEs that were not reported is obviously very costly and even unrealistic. Instead, systems that employ machine learning methods are more efficient – and perhaps not less effective – by learning patterns from patients who have ADEs encoded in their medical history and detecting patients that have suffered an ADE, yet in whose records it has not, but should have, been encoded. Employing such systems could result in improved reporting of ADEs and, in the long run, potentially reduce the unsafe use of drugs.

There are two parallel approaches for such systems to facilitate the reporting of ADEs: (1) to prospectively predict whether there are ADEs to report in new health records; and (2) to retrospectively detect ADEs that are missing in earlier health records. Both approaches require the underlying models to include information from patients’ medical history, given that the first approach needs to use patient history to train the predictive models. The problem thus boils down to making meaningful use of patient history information in EHRs in order to detect ADEs. This problem has previously been studied from different perspectives, such as how best to represent clinical events¹²,¹³. In this study, we focus on the temporal aspect, i.e., which time periods should be considered when collecting clinical events for predicting ADEs. This aspect is expected to have a high impact on the resulting predictive performance, as the probability of an adverse event of a drug occurring at a certain point in time is highly dependent on when the drug was taken, e.g., an ADE will typically not occur for the first time years after a drug has been taken. On the one hand,
considering longer time periods may lead to the ADE signal drowning in the volumes of collected information, while, on the other hand, shorter time periods may exclude information that is crucial for making accurate predictions.

In previous research, a few studies have touched upon or analyzed temporal relationships in EHRs for detecting ADEs. Norén et al. monitor clinical events that have occurred before and after a drug is prescribed to look for abnormalities as signals indicating ADEs\textsuperscript{14}. Their study is, however, in a different genre as it does not involve a learning procedure. At the same time, Jin et al. transform the temporal problem onto a cross-section problem by defining hazard period, effect period and reference period after a drug is prescribed\textsuperscript{15}. In a recent study, Eriksson et al. extract clinical notes and structured prescription data between drug introduction and discontinuation and then filter ADEs based on time stamp inconsistency between structured data and notes\textsuperscript{16}. However, all of these studies mainly look at the temporality on a rather detailed level, such as between specific events or predefined periods; how to handle the temporality in EHRs in general still remains unanswered. In this study, the scope is set to detecting ADEs that should be reported in EHRs by employing machine learning methods to learn patterns using clinical events, and hence the essence of the problem is how to model the events that occurred at different time points in the patient history. To the best of our knowledge, this is the first study to explore ways of handling the temporality of patient history prior to an ADE. The aim of this study is therefore two-fold: (1) to investigate various ways of handling the temporality of clinical events; (2) to explore the importance of clinical events at different time points for detecting ADEs.

**Materials and Methods**

In this section, we first introduce the strategies of handling temporality that are proposed in this study. Then, we describe a series of empirical experiments that are conducted to evaluate the proposed strategies, starting by describing the data source and subsequently the details of each experiment.

**- Strategies to handle temporality**

Three strategies – *bag of events*, *bag of binned events*, and *bag of weighted events* – to handle the temporality of clinical events in patients’ medical history for the detection of ADEs were proposed and evaluated in this study. In each of the strategies, only clinical events that occurred prior to the target ADE were included. A toy example of handling the temporality of one drug, one diagnosis and one clinical measurement in a time period of three days is demonstrated in Figure 1.

Here, let \( x \) denote each unique event in the whole EHR database, \( d \) denote the number of days prior to the occurrence of the target ADE, \( d \leq D \), and \( n_d \) denote the number of times \( x \) occurred in day \( d \).

- **Bag of Events (BE)** This strategy counts the number of occurrences of event \( x \) within \( D \) days. In this case, the value of feature \( x \) is \( \sum_{d=1}^{D} n_d \).

- **Bag of Binned Events (BBE)** This strategy counts the number of occurrences of event \( x \) in each day within \( D \) and represents \( x \) as \( x_1, x_2, ..., x_D \). The value of the corresponding feature here is \( n_1, n_2, ..., n_D \).

- **Bag of Weighted Events (BWE)** This strategy assigns different weights to event \( x \) that occurred at different days \( d \) and takes into account the weights when counting the number of occurrences of \( x \). The proportional weights are assigned according to the time distance between the event and the target ADE: events that occurred further in time from the target ADE receive proportionally less weight. In this case, the time distance between \( x \) at day \( d \) and the target ADE is \( d \), and hence the value of feature \( x \) is \( \sum_{d=1}^{D} (n_d / d) \).

**- Data source**

In this study, data was extracted from the Stockholm EPR Corpus\textsuperscript{17}, which contains health records of around 700,000 (anonymized) patients over a two-year period (2009-2010), obtained from Karolinska University Hospital in Stockholm, Sweden. This database contains large amounts of diagnosis information, drug administrations, clinical measurements and clinical notes in free-text. In this study, we used only the structured data, i.e., diagnoses, drugs and clinical...
Figure 1: Strategies to handle temporality of three clinical events during three days

measurements, as features. In the Stockholm EPR Corpus, diagnoses are encoded by the International Statistical Classification of Diseases and Related Health Problems, 10th Edition (ICD-10) and drugs are encoded by the Anatomical Therapeutic Chemical Classification System (ATC). This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm) with permission number 2012/834-31/5.

The targeted use case in this study is to detect patients who should, but do not, have a specific ADE reported in their health records by retrospectively analyzing clinical events in their medical history. Among the diagnosis codes that indicate ADEs\(^\text{\ref{ADE}}\), category A.1 (a drug-related causation was noted in the diagnosis code) and category A.2 (a drug- or other substance-related causation was noted in the diagnosis code) were considered in this study. We selected the 14 most frequent A.1 and A.2 codes in the Stockholm EPR Corpus, and thus created 14 datasets, where the existence of the diagnosis code indicating a particular ADE was used as the class label in each dataset. The classification task is hence binary: positive or negative towards a specific ADE. Both inpatients and outpatients are included in this study, where the positive examples are patients whom have been assigned an ADE-specific diagnosis code and the negative examples are patients whom have been assigned a similar diagnosis code to the diagnosis code indicating ADE. Here, similarity is defined as two codes sharing the same first three levels of the ICD-10 concept hierarchy. For instance, if the positive examples are patients with diagnosis code \(G44.4\) (drug-induced headache), the negative examples are patients with any diagnosis code starting with \(G44\) (other headache syndromes), but not \(G44.4\). Table 1 lists the basic descriptions of each dataset, including the diagnosis code that indicates the corresponding ADE (dataset name), the description of this code and the number of positive and negative examples.

Around 10,000 unique ICD-10 diagnosis codes, 1,500 unique ATC codes and 730 unique clinical measurements exist in the Stockholm EPR Corpus. Since most clinical events only occurred to a small group of patients, the datasets where clinical events are used to describe each patient are consequently both high-dimensional and very sparse, i.e., for a single observation (or example) the vast majority of the features have a value of 0. The employed algorithm for generating predictive models, i.e., the random forest algorithm\(^\text{\ref{random_forest}}\), is rather efficient in handling high-dimensional data, as only a small random sample of the available features is considered when determining the best way of partitioning the training examples during tree growth. However, for highly sparse data, it is not unlikely that all sampled features are uninformative, i.e., lead to no separation. Unless specifically handled, the tree growth will terminate prematurely and lead to an overall low predictive performance. Rather than employing any of the more sophisticated approaches to handle this\(^\text{\ref{sparse_data}}\), we here employed the quite straightforward approach of removing features that are more sparse than 99%, i.e., the ones for which non-zero values were observed in fewer than 1% of the patients; for the datasets with fewer than one hundred observations, features with only one non-zero value were also removed. The motivation behind this is simply that for features with very few non-zero values, the impact of applying different strategies to handle temporality will be negligible, even though some of these features might be valuable indicators.
Table 1: Dataset description

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Corresponding diagnosis code description</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>D642</td>
<td>Secondary sideroblastic anemia due to drugs and toxins</td>
<td>113</td>
<td>4234</td>
</tr>
<tr>
<td>G240</td>
<td>Drug-induced dystonia</td>
<td>16</td>
<td>44</td>
</tr>
<tr>
<td>G444</td>
<td>Drug-induced headache, not elsewhere classified</td>
<td>31</td>
<td>1102</td>
</tr>
<tr>
<td>G620</td>
<td>Drug-induced polyneuropathy</td>
<td>19</td>
<td>367</td>
</tr>
<tr>
<td>I952</td>
<td>Hypotension due to drugs</td>
<td>38</td>
<td>480</td>
</tr>
<tr>
<td>L270</td>
<td>Generalized skin eruption due to drugs and medicaments</td>
<td>174</td>
<td>291</td>
</tr>
<tr>
<td>L271</td>
<td>Localized skin eruption due to drugs and medicaments</td>
<td>58</td>
<td>407</td>
</tr>
<tr>
<td>O355</td>
<td>Maternal care for (suspected) damage to fetus by drugs</td>
<td>334</td>
<td>373</td>
</tr>
<tr>
<td>T782</td>
<td>Adverse effects: anaphylactic shock, unspecified</td>
<td>136</td>
<td>1467</td>
</tr>
<tr>
<td>T783</td>
<td>Adverse effects: angioneurotic oedema</td>
<td>147</td>
<td>1448</td>
</tr>
<tr>
<td>T784</td>
<td>Adverse effects: allergy, unspecified</td>
<td>984</td>
<td>612</td>
</tr>
<tr>
<td>T808</td>
<td>Other complications following infusion, transfusion and therapeutic injection</td>
<td>353</td>
<td>59</td>
</tr>
<tr>
<td>T886</td>
<td>Anaphylactic shock due to correct drug or medicament properly administered</td>
<td>53</td>
<td>607</td>
</tr>
<tr>
<td>T887</td>
<td>Unspecified adverse effect of drug or medicament</td>
<td>472</td>
<td>277</td>
</tr>
</tbody>
</table>

In this study, we defined the following 12 thresholds from patient history: 1, 2, 3, 4, 5, 6, 7, 14, 21, 30, 60, 90 days before the target ADE, where, for each threshold \( n \), clinical events that occurred \( n \) days before the target ADE are studied. The choice was made with the assumption that more clinical events would have occurred in the days closer to the ADE and the difference between days closer to the ADE has a higher impact on detecting ADEs, especially on the ones that are dose-independent. We therefore studied clinical events that occurred in each day of the first week, followed by each week in the first month, and finally every month up to 3 months (90 days) before the occurrence of the target ADE. In the *bag of binned events* strategy, clinical events are binned within a time threshold. For example, all events that occurred one day before ADE are binned in day 1, and all events that occurred from 60 days to 90 days before the ADE are binned in day 90. Here, the same event that occurred at different times is treated as different events, the number of features in *bag of binned events* is hence much higher than the other two strategies. To reduce the number of potential models, we do not distinguish between the events that occurred in different days within, e.g., the third month before the ADE. The average number of features and sparsity over the 14 datasets, before and after removing the extremely sparse features at each time threshold, are illustrated in Figure 2.

![Figure 2: Average number of features and sparsity before and after removing extremely sparse features](image)

**Experiments**

In this study, a series of experiments was conducted to evaluate the strategies to handle temporality of clinical events in patient history and explore the impact of using different lengths of patient history on detecting ADEs. In the first experiment, the three strategies — *bag of events*, *bag of binned events* and *bag of weighted events* — were compared by employing the random forest\(^9\) learning algorithm to features (clinical events) generated by each of them respectively.
When more than two competing models are compared, the Friedman test was employed for statistical testing of the null hypothesis that all models perform equally, followed by a post-hoc test using the Bergman-Hommel procedure. The choice of random forest was made due to its reputation of achieving high accuracy, its ability to handle high-dimensional data, as well as the possibility of obtaining estimates of variable importance. The algorithm constructs an ensemble of decision trees, which together vote for what class label to assign to an example that is to be classified. Each tree in the forest is built from a bootstrap replicate of the original instances, and a subset of all features is sampled at each node when building the tree, in both cases to increase diversity among the trees. When the number of trees in the forest increases, the probability that a majority of trees makes an error decreases, given that the trees perform better than random and that the errors are made independently. Although this condition can only be guaranteed in theory, the algorithm has often been shown in practice to result in state-of-the-art predictive performance. In this study, we used random forest with 500 trees. In all experiments, models were built and evaluated using stratified 5-fold cross-validation with two iterations, where the original class distribution was retained in each fold.

The considered performance metrics are accuracy, area under the ROC curve (AUC), precision, recall, F-score and area under the precision-recall curve (AUPRC). Accuracy corresponds to the percentage of correctly classified instances. AUC depicts the performance of a model without regard to class distribution or error costs by estimating the probability that a model ranks a randomly chosen positive instance ahead of a negative one. Precision measures the fraction of true positives among all the predicted positives, while recall, also known as sensitivity, measures the fraction of true positives among all the positives in the gold standard. In the case of detecting ADEs, high precision means that the algorithm is able to detect more true ADEs than false ones, while high recall means that the algorithm is able to detect most true ADEs. F-score describes the harmony between precision and recall by calculating $2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$. Only both high precision and high recall can yield a high F-score. AUPRC depicts the probability that precision is higher than recall for each recall threshold. It is considered to be a more careful measurement compared to AUC, since high AUC often leads to high AUPRC, but not necessarily the other way around.

In a subsequent set of experiments, variable importance generated with random forest using bag of binned events was analyzed in order to obtain a deeper understanding of in which time period clinical events are more informative. Variable importance can be estimated in different ways. In this study, Gini importance was chosen as the variable was analyzed in order to obtain a deeper understanding of in which time period clinical events are more informative.

Each tree in the forest is built from a bootstrap replicate of the original instances, and a subset of all features is sampled at each node when building the tree, in both cases to increase diversity among the trees. When the number of trees in the forest increases, the probability that a majority of trees makes an error decreases, given that the trees perform better than random and that the errors are made independently. Although this condition can only be guaranteed in theory, the algorithm has often been shown in practice to result in state-of-the-art predictive performance. In this study, we used random forest with 500 trees. In all experiments, models were built and evaluated using stratified 5-fold cross-validation with two iterations, where the original class distribution was retained in each fold.

Friedman tests were conducted to compare the three strategies at each time threshold over the 14 datasets (p-values are shown in Table 2). P-values are not available for time threshold $l$ day since the three strategies are identical for
clinical events that occur one day before the ADE. From this table we can see that the p-values generally decrease with more days of patient history, and the null hypothesis that the three strategies perform equally well is rejected with most evaluation metrics towards the end, i.e., using 90 days of patient history. This indicates that when we include more days of patient history in our predictive models, the impact of how temporality is handled increases. Note that although there seems to be a performance drop from 60 days to 90 days in Figure 3, if we take a closer look at the performance on each dataset (e.g., Figure 4 for bag of weighted events), we can see that for most datasets, the predictive performance improves from 60 days to 90 days, except for a big drop on two (G620 and L271), which explains most of the overall performance drop observed in Figure 3.

Table 2: P-values of statistical significance of differences among the three strategies with each time threshold

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>14</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.54</td>
<td>0.84</td>
<td>0.66</td>
<td>0.37</td>
<td>0.56</td>
<td>0.3</td>
<td>0.048</td>
<td>0.74</td>
<td>0.12</td>
<td>0.07</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>0.93</td>
<td>0.81</td>
<td>0.22</td>
<td>0.22</td>
<td>0.61</td>
<td>0.17</td>
<td>0.26</td>
<td>0.11</td>
<td>0.07</td>
<td>0.002</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.75</td>
<td>0.75</td>
<td>0.07</td>
<td>0.9</td>
<td>0.61</td>
<td>0.08</td>
<td>0.61</td>
<td>0.08</td>
<td>0.07</td>
<td>0.003</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.93</td>
<td>0.42</td>
<td>0.61</td>
<td>0.02</td>
<td>0.1</td>
<td>0.1</td>
<td>0.04</td>
<td>0.59</td>
<td>0.04</td>
<td>0.029</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.83</td>
<td>0.84</td>
<td>0.72</td>
<td>0.84</td>
<td>0.27</td>
<td>0.61</td>
<td>0.9</td>
<td>0.53</td>
<td>0.54</td>
<td>0.57</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>F-score</td>
<td>0.64</td>
<td>0.49</td>
<td>0.16</td>
<td>0.12</td>
<td>0.18</td>
<td>0.7</td>
<td>0.64</td>
<td>0.13</td>
<td>0.04</td>
<td>0.27</td>
<td>0.025</td>
<td></td>
</tr>
</tbody>
</table>

Given these results, a post-hoc analysis comparing the strategies pairwise was carried out with 90 days as threshold; the results are shown in Table 3. The left half of the table shows the average rank – the lower the rank, the better the performance – among the three strategies, and we can see that bag of weighted events is consistently better than the other two. The pairwise p-values in the right half of the table indicate that bag of weighted events is significantly better than bag of events for AUPRC, precision and F-score; bag of weighted events is significantly better than bag of binned events for AUC and AUPRC; and bag of binned events is significantly better than bag of events for AUC.

- Impact of temporality on specific adverse drug events

The predictive performance of the best observed strategy – bag of weighted events – for each specific ADE is presented in Figure 4. As described in the previous section, the predictive performance overall increases on a small scale towards
Table 3: Average ranks and p-values from the post-hoc analysis of Friedman tests with 90 days as threshold

<table>
<thead>
<tr>
<th>Metric</th>
<th>BE</th>
<th>BBE</th>
<th>BWE</th>
<th>BBE vs. BWE</th>
<th>BE vs. BWE</th>
<th>BBE vs. BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>2.14</td>
<td>2.29</td>
<td>1.57</td>
<td>0.176</td>
<td>0.176</td>
<td>0.705</td>
</tr>
<tr>
<td>AUC</td>
<td>1.96</td>
<td>2.71</td>
<td>1.32</td>
<td>0.0007</td>
<td>0.089</td>
<td>0.047</td>
</tr>
<tr>
<td>AUPRC</td>
<td>2.36</td>
<td>2.43</td>
<td>1.21</td>
<td>0.0004</td>
<td>0.004</td>
<td>0.85</td>
</tr>
<tr>
<td>Precision</td>
<td>2.57</td>
<td>1.89</td>
<td>1.54</td>
<td>0.34</td>
<td>0.018</td>
<td>0.07</td>
</tr>
<tr>
<td>Recall</td>
<td>2.28</td>
<td>1.96</td>
<td>1.75</td>
<td>0.57</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>F-score</td>
<td>2.5</td>
<td>1.96</td>
<td>1.54</td>
<td>0.26</td>
<td>0.032</td>
<td>0.156</td>
</tr>
</tbody>
</table>

using more days of patient history, which is especially clear for D642, T808 and T887. Even with an obvious drop at the end on e.g. G620, the predictive performance with 90 days is still higher than it is with 1 day. Apparently, the choice of threshold has different degrees of impact on different ADEs.

**Figure 4: Predictive performance of random forest with bag of weighted events for each ADE**

To obtain a deeper understanding of the contribution of clinical events from different days of patient history to the predictive performance, variable importance analysis was conducted using bag of binned events, where they are treated as different variables. Figure 5 shows the relative importance of clinical events occurring in different days of patient history for detecting each specific ADE. For most ADEs, it is clear that clinical events occurring in the late stage of patient history, e.g., just one day before the ADE, is of highest importance; for some ADEs, such as D642 and G620, clinical events occurring in a much earlier stage of patient history are relatively more important.

**Discussion**

For the task of detecting ADEs using clinical events in EHRs, the predictive performance with the three strategies differs from each other within a fairly small range, even though there are significant differences observed. This can be explained by the difficulty of the task itself. Here, in each dataset, the positive examples are patients who have a specific ADE, while the negative examples are patients who have a disease that is in the same family as the ADE (they share the first three levels of ICD-10). In many cases, these patients are very similar to each other in terms of having similar drugs and clinical measurements; therefore it is difficult for the learning algorithm to distinguish between them. Hence it is difficult to observe big changes in predictive performance with different strategies since they are all bounded with the similar clinical events these patients share.
Based on our results, the random forest built with \textit{bag of events} has the worst predictive performance, followed by \textit{bag of binned events}, and both of them are, in most cases, outperformed by the random forest built with \textit{bag of weighted events}. This finding is consistent with the granularity of each of the strategies: the model with \textit{bag of events} completely ignores temporality, and hence is the crudest model; although the model with \textit{bag of binned events} treats events from different time periods as different features, such a representation also results in increased dimensionality and sparsity that have a negative impact on the predictive performance; and, finally, the model with \textit{bag of weighted events} tackled the problems with the other two strategies – temporality, dimensionality and sparsity – by aggregating the events from different time periods according to their time distance from the target ADE.

The predictive performance and variable importance for each specific ADE to some extent reflect the different nature of each ADE. Adverse drug events are typically divided into two types: dose-dependent or not. The former is related to the accumulation of toxics from drugs or medications; therefore, when predicting ADEs of this type, we would intuitively expect that using events from longer patient history contributes to improved predictive performance compared to using only events from the most recent days. This assumption is supported by the results on predicting \textit{D642 – drug induced anemia}, where the predictive performance improves monotonically with the use of more days of patient history (see Figure 4 – D642) and the clinical events from seven days prior to the ADE are much more important compared to one day (see Figure 5 – D642). Another example is \textit{G620 – drug induced polyneuropathy}, which is an ADE known to be caused by drugs used in the treatment of very severe or chronic diseases like cancer or tuberculosis. Such patients are normally exposed for a long period until the side effect is observed. Here, again, a similar performance trend is observed to support our hypothesis (see Figure 4 – G620 and Figure 5 – G620). On the other hand, the second type of ADE is related mainly to idiosyncratic or immunological nature, such as an allergic reaction, which indicates that most likely, with a very marginal assumption, they are more or less instant side effects from taking drugs or medications. The results on predicting \textit{T886 – anaphylactic shock due to correct drug or medication properly administered} is a good example that agrees with this assumption. The predictive performance almost remains the same as using events from only one day before the ADE (see Figure 4 – T886) and the events from one day before are absolutely the most important ones (see Figure 5 – T886).

However, the assumptions that are made above are not always reflected by our results in this study, which can be explained by several reasons. First, only clinical events from the structured EHRs are used as features here, while ADEs are often described in the notes rather than reported as diagnoses (given that ADEs are heavily under-reported as we described in the introduction). Especially for patients who suffer from severe diseases, ADEs become less
important to report. Therefore, using only the structured clinical events can hardly capture a holistic and precise picture for detecting ADEs. Second, in the real clinical setting, where the EHR data is collected, there is a lack of controls on when clinical events, especially diagnosis codes, are reported in the database. Sometimes events that occurred weeks ago are reported together with the ones that occurred one day ago when the patients are discharged. This, unfortunately, results in noise in the training data and also inaccurate weights based on the time distance from the target ADE. Third, diagnosis codes, i.e., the ICD-10 codes, are themselves heterogeneous. For example, T782 - T784 cover not only ADEs but also adverse reactions to substances that are not drugs.

One limitation of this study concerns the fact that the events are represented as a bag in all of the strategies, which neglects the temporality between the events in the same day or period. This has a potential impact on the predictive performance, especially for predicting ADEs that are not dose-dependent, since the temporal relationship between events within a short period is important to capture when the ADEs are mostly instant effects. To take this into account, techniques that enable mining event sequences would be useful. In addition, the use of ICD-10 codes to select patients with and without ADEs should be proceeded with caution due to the fact that they are coded also for billing purposes and other reasons. For future work, it would also be interesting to see the impact of the proposed strategies on the task of distinguishing patients who have ADEs from patients who do not, i.e., the negative examples would be randomly selected from the population. The way of assigning weights in bag of weighted events is fairly crude and harsh; a follow-up study could focus on exploring alternative ways of assigning weights to clinical events that occur at different time points. To improve the predictive performance, including free-text clinical notes in the models and handling the temporality of information in notes are indeed worth investigating.

Conclusions

It is advantageous to use longitudinal data from electronic health records for the detection of adverse drug events. This study proposed and evaluated three strategies to handle temporality of clinical events: drugs, diagnosis and clinical measurements. These strategies differ from each other in how they take into account the clinical events that occur in different days of patient history: bag of events counts the number of times that an event occurred without regard to when it occurred; bag of binned events separates the patient history into predefined bins and then counts the number of times that an event occurred in each bin; and bag of weighted events counts the weighted number of times that an event occurred, where the weights are assigned proportionally according to time distance between the event and the target adverse drug event. Based on our empirical investigation, the bag of weighted events strategy yields the best predictive performance with the considered metrics. Here, we conclude that, in general, using longer patient history leads to improved predictive performance, and the temporality of clinical events matters more when using more days of patient history.

Acknowledgements

This work was supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection, ref. no. IIS11-0053.

References


Using a Clinical Knowledge Base to Assess Comorbidity Interrelatedness Among Patients with Multiple Chronic Conditions

Donna M. Zulman, MD, MS,1,2 Susana B. Martins, MD, MSc,3 Yan Liu, PhD,4 Samson W. Tu, MS,5 Brian B. Hoffman, MD,6,7 Steven M. Asch, MD,1,2 Mary K. Goldstein, MD, MS3,8

1 Center for Innovation to Implementation, VA Palo Alto Health Care System, Menlo Park, CA; 2 Division of General Medical Disciplines, Stanford University, Stanford, CA; 3 Geriatrics Research Education and Clinical Center (GRECC), VA Palo Alto Health Care System, Palo Alto, CA; 4 Department of Biomedical, Industrial, and Human Factors Engineering, Wright State University, Dayton, OH; 5 Biomedical Informatics Research (BMIR), Stanford University School of Medicine, Stanford, CA; 6 VA Boston Health Care System, West Roxbury Division, Boston, MA; 7 Harvard Medical School, Boston, MA; 8 Center for Primary Care and Outcomes Research (PCOR), Stanford University School of Medicine, Stanford, CA

Abstract

Decision support tools increasingly integrate clinical knowledge such as medication indications and contraindications with electronic health record (EHR) data to support clinical care and patient safety. The availability of this encoded information and patient data provides an opportunity to develop measures of clinical decision complexity that may be of value for quality improvement and research efforts. We investigated the feasibility of using encoded clinical knowledge and EHR data to develop a measure of comorbidity interrelatedness (the degree to which patients’ co-occurring conditions interact to generate clinical complexity). Using a common clinical scenario—decisions about blood pressure medications in patients with hypertension—we quantified comorbidity interrelatedness by calculating the number of indications and contraindications to blood pressure medications that are generated by patients’ comorbidities (e.g., diabetes, gout, depression). We examined properties of comorbidity interrelatedness using data from a decision support system for hypertension in the Veterans Affairs Health Care System.

Introduction

The combined forces of an aging population and rising rates of chronic disease have resulted in a rapid increase in the number of adults experiencing multimorbidity, or multiple chronic conditions (MCCs).1,2 The presence of MCCs generates a number of challenges for health care delivery. As a patient’s number of chronic conditions increases, there is an associated rise in physician visits, prescriptions, hospitalizations, and home health services.1,3-6 Patients with MCCs are also at greater risk for mortality and more rapid functional decline.1,7-10

For clinicians, patients with MCCs present unique challenges because of the need to address multiple health issues, often during a brief clinic visit and in the face of competing psychosocial demands.11-13 Patients and their providers may also have different priorities regarding which of a patient’s conditions are most important.14 Traditional multimorbidity measures such as the Charlson score15,16 often incorporate disease burden, but may not capture the full complexity involved when caring for patients with MCCs. Newer conceptual frameworks for multimorbidity have started to incorporate biological, socioeconomic, cultural, environmental, and behavioral health determinants.17 We previously proposed a framework that also includes comorbidity interrelatedness, or the degree to which patients’ co-occurring conditions interact to generate clinical decision-making complexity.18 In order for these frameworks to influence clinical practice, however, there is a need to operationalize constructs such as comorbidity interrelatedness so that they may be applied to quality improvement, performance evaluation, and research efforts.

Knowledge-based decision support tools integrate clinical knowledge (e.g., information about potential medication interactions) with electronic health record (EHR) patient data to provide clinicians with tailored messages about optimal therapy choices for specific patients. Such tools may also provide an opportunity to quantify measures of
clinical complexity such as comorbidity interrelatedness. The objective of this study was to test the feasibility of harnessing encoded knowledge and EHR data to calculate a comorbidity interrelatedness score, and to explore this domain’s properties in a population. Using a common clinical scenario—decisions about blood pressure medications in patients with a hypertension diagnosis—we determined the degree to which comorbidities influence medication decisions in patients with MCCs, and examined variation in this phenomenon across a population.

Methods

ATHENA-HTN

In order to test the feasibility of constructing a comorbidity interrelatedness score, we obtained encoded knowledge and patient-level EHR data from a study of a clinical decision support (CDS) module, ATHENA-HTN. ATHENA-HTN presents detailed patient-specific evidence-based recommendations for hypertension management to primary care providers at point-of-care.19-23 ATHENA-HTN includes a knowledge base (KB) that models hypertension knowledge in computer-interpretable formats. A guideline interpreter (execution engine) processes EHR data (e.g., diagnoses, medications, blood pressure values, and laboratory test results) and generates automated patient-specific recommendations for medications using information from the KB (Figure 1).19 The KB was created with Protégé, a widely used open-source knowledge acquisition software using the EON model24-26 In developing the KB, clinical knowledge from hypertension guidelines and other sources was reviewed and discussed with subject matter experts then entered manually into the Protégé KB. Medication indications and contraindications derived from published guidelines, pharmacology textbooks, and subject matter experts.27-29 The current project took advantage of this encoded clinical knowledge to calculate comorbidity interrelatedness scores for patients enrolled in the ATHENA-HTN study.

Figure 1. Derivation of a comorbidity interrelatedness score that integrates patients’ electronic health record data with encoded clinical knowledge from a hypertension decision support tool to determine the degree to which an individual’s chronic conditions influence clinical decisions about hypertension.

The current analyses use a dataset from a clinical implementation study of ATHENA-HTN in which patient-level clinical data (e.g., comorbidities) were linked to encoded knowledge about comorbidity-drug class interactions.30, 31 The study population for the ATHENA-HTN trial comprised Veterans Affairs (VA) patients with an ICD-9-based diagnosis of hypertension who had one or more primary care clinic visits with an enrolled provider during the study period. Because ATHENA-HTN is designed to provide treatment recommendations for patients with primary
hypertension who are suitable for standard hypertension guidelines, the study excluded patients who could be identified as having secondary hypertension, those on 5 or more antihypertensive agents, and those who had other circumstances such that standard guidelines would not apply (for example, patients who were under 21 years; patients who were pregnant; patients with spinal cord injury, narcolepsy, ascites, hypertrophic cardiomyopathy, or a home-bound clinical state; and patients with active prescriptions for medication(s) that indicated standard guidelines would not apply such as transplant drugs). The study using ATHENA-HTN was approved by the Human Subjects Panel, Stanford University, and the Research and Development (R&D) Committee, VA Palo Alto Health Care System.

**Patient Selection for Feasibility Study**

We obtained data for 5,997 individuals in the ATHENA-HTN study database who had a blood pressure reading above goal at their first clinic visit during the study period. Blood pressure goals were defined using VA/Department of Defense guidelines at that time (i.e., < 140/90mmHg for patients without diabetes and < 140/80mmHg for patients with diabetes). We focused on patients with elevated blood pressure because they represent individuals who warrant consideration for medication modification or intensification. In addition, patients with blood pressure above established goals are frequently flagged by EHR applications that aim to improve hypertension management quality of care and performance.

**Comorbidity Interrelatedness Score for Patients with Hypertension**

In developing a comorbidity interrelatedness score, we first identified comorbid conditions that commonly affect choice of hypertension medications by generating an indication or contraindication to a medication class. We considered a comorbid condition to be hypertension-related if published clinical evidence or guidelines describe interactions between the condition and one or more major anti-hypertensive drug classes. For example, clinical guidelines recommend that patients with diabetes receive an ACE-inhibitor (or an angiotensin receptor blocker) to decrease risk of cardiovascular events and progressive kidney dysfunction. Patients with osteoporosis, meanwhile, may benefit from a thiazide because these medications promote calcium reabsorption by the kidney. For patients with gout, however, thiazides should be avoided as they can increase hyperuricemia and result in a gout flare, particularly if the gout is poorly controlled. We identified the following conditions: Diabetes, Coronary Artery Disease, Benign Prostatic Hyperplasia, Depression, Obstructive Pulmonary Disease (Non-Bronchospastic), Gout, Cerebrovascular Disease, Atrial Fibrillation, Myocardial Infarction History, Heart Failure, Bronchospastic Disease, Angina, Atrial Tachycardia, Urinary Incontinence, Osteoporosis, Chronic Kidney Disease, Sinoatrial Node Dysfunction (no pacemaker), Migraine, Heart Block (1st degree), Hyperthyroidism, Heart Block (2nd/3rd degree), and Raynaud’s.

The comorbidity interrelatedness score incorporates information about the number of drug classes that are indicated or contraindicated (or both) based on a patient’s chronic conditions. We included six anti-hypertensive drug classes in the comorbidity interrelatedness measure: thiazides, ACE-inhibitors, angiotensin-receptor blockers, cardioselective beta-receptor antagonists, dihydropyridine (DHP) calcium-channel blockers, and non-DHP calcium-channel blockers. These drug classes were selected because they are commonly prescribed for hypertension. We did not include alpha-blockers, loop diuretics, or potassium-sparing diuretics in the measure because their primary indication was often a condition other than hypertension (e.g., prostatic hyperplasia, heart failure, and liver disease, respectively). Prescriptions for non-cardioselective beta-receptor antagonists, central adrenergic anti-hypertensives, direct vasodilators, and peripheral adrenergic anti-hypertensives were present for 2% or fewer subjects and were therefore excluded from the measure as well.

Figure 2 illustrates the number of anti-hypertensive drug classes that are either indicated or contraindicated in the setting of specific hypertension-related comorbidities. For each individual, we calculated a comorbidity interrelatedness score equal to the sum of all comorbidity-related drug class indications and contraindications that were generated by his or her specific conditions. For example, a patient with diabetes, gout, and depression would receive a score of 7 because gout and depression each generate a relative contraindication for a single anti-hypertensive drug class, and diabetes generates indications for five different anti-hypertensive drug classes. We also calculated the number of anti-hypertensive classes to which each individual had both an indication and a contraindication (a circumstance that generates an extra level of complexity because a provider needs to weigh the potential benefits of the medication for Condition A against the potential risks for Condition B). For the patient above, diabetes, gout, and depression generate an indication and a contraindication for thiazides and cardioselective beta-receptor antagonists (Figure 2). To examine patterns of complexity in the population, we calculated the mean
(SD) comorbidity interrelatedness score, the score distribution, and the mean (SD) number of anti-hypertension classes to which individuals had both an indication and a contraindication, for all individuals with 0 through 6 hypertension-related comorbidities.

<table>
<thead>
<tr>
<th>Anti-Hypertensives</th>
<th>Comorbid Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE-Inhibitor</td>
<td>Angina</td>
</tr>
<tr>
<td>Angiotensin</td>
<td>Atrial Fibrillation</td>
</tr>
<tr>
<td>Receptor Blocker</td>
<td>Atrial Tachycardia</td>
</tr>
<tr>
<td>Beta Receptor</td>
<td>Benign Prostatic</td>
</tr>
<tr>
<td>Antagonist</td>
<td>Hyperplasia</td>
</tr>
<tr>
<td>(cardioselective)</td>
<td>(on alpha blocker)</td>
</tr>
<tr>
<td>Calcium Channel</td>
<td>Bronchospastic</td>
</tr>
<tr>
<td>Blocker (DHP)</td>
<td>Disease</td>
</tr>
<tr>
<td>Calcium Channel</td>
<td>Cerebrovascular</td>
</tr>
<tr>
<td>Blocker (Non-DHP)</td>
<td>Disease</td>
</tr>
<tr>
<td>Thiazide</td>
<td>Chronic Kidney</td>
</tr>
<tr>
<td></td>
<td>Disease</td>
</tr>
</tbody>
</table>

*Figure 2:* Indications and contraindications to anti-hypertensive drug classes arising from comorbid chronic conditions.

**Results**

Table 1 describes sociodemographic and clinical characteristics of patients from the ATHENA-HTN trial who were included in this feasibility study. The mean (SD) age of patients was 71 (11), and the vast majority were male (97%). The mean (SD) number of prescribed anti-hypertensive classes was 1.6 (1.1), but 22% of the patients had three or more prescriptions. The most common comorbidities affecting anti-hypertensive class choice were diabetes (42%), coronary artery disease (33%), depression (23%), obstructive pulmonary disease (16%), gout (11%), and cerebrovascular disease (10%).
Table 1: Patient Characteristics

<table>
<thead>
<tr>
<th></th>
<th>n (%)</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age in years, mean (SD)</strong></td>
<td>—</td>
<td>71 (11)</td>
</tr>
<tr>
<td>Age &gt; 65 years</td>
<td>4,061 (68)</td>
<td></td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td>5,829 (97)</td>
<td></td>
</tr>
<tr>
<td><strong>Systolic blood pressure, mmHg</strong></td>
<td>—</td>
<td>148 (14)</td>
</tr>
<tr>
<td><strong>Diastolic blood pressure, mmHg</strong></td>
<td>—</td>
<td>79 (10)</td>
</tr>
<tr>
<td><strong>Prescribed Anti-Hypertensive Classes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACE-Inhibitor</td>
<td>2,827 (50)</td>
<td></td>
</tr>
<tr>
<td>Angiotensin-Receptor Blocker</td>
<td>811 (14)</td>
<td></td>
</tr>
<tr>
<td>Beta-Receptor Antagonist (cardioselective)</td>
<td>2,431 (43)</td>
<td></td>
</tr>
<tr>
<td>Calcium-Channel Blocker (dihydropyridine)</td>
<td>1,271 (22)</td>
<td></td>
</tr>
<tr>
<td>Calcium-Channel Blocker (non-dihydropyridine)</td>
<td>370 (7)</td>
<td></td>
</tr>
<tr>
<td>Thiazide</td>
<td>2,068 (36)</td>
<td></td>
</tr>
<tr>
<td><strong>Number of Prescribed Anti-Hypertensive Classes</strong></td>
<td>—</td>
<td>1.6 (1.1)</td>
</tr>
<tr>
<td>0</td>
<td>990 (17)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1,748 (29)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1,974 (33)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1,062 (18)</td>
<td></td>
</tr>
<tr>
<td>4+</td>
<td>223 (4)</td>
<td></td>
</tr>
<tr>
<td><strong>Conditions that Affect Anti-Hypertensive Class Choice</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angina</td>
<td>210 (4)</td>
<td></td>
</tr>
<tr>
<td>Atrial Fibrillation</td>
<td>536 (9)</td>
<td></td>
</tr>
<tr>
<td>Atrial Tachycardia</td>
<td>130 (2)</td>
<td></td>
</tr>
<tr>
<td>Benign Prostatic Hyperplasia (and on alpha blocker)</td>
<td>535 (9)</td>
<td></td>
</tr>
<tr>
<td>Bronchospastic Disease</td>
<td>326 (5)</td>
<td></td>
</tr>
<tr>
<td>Cerebrovascular Disease</td>
<td>593 (10)</td>
<td></td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>103 (2)</td>
<td></td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td>1,955 (33)</td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>1,395 (23)</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>2,532 (42)</td>
<td></td>
</tr>
<tr>
<td>Gout</td>
<td>676 (11)</td>
<td></td>
</tr>
<tr>
<td>Heart Block (1st degree)</td>
<td>64 (1)</td>
<td></td>
</tr>
<tr>
<td>Heart Block (2nd/3rd degree)</td>
<td>21 (&lt;1)</td>
<td></td>
</tr>
<tr>
<td>Heart Failure</td>
<td>357 (6)</td>
<td></td>
</tr>
<tr>
<td>Hyperthyroidism</td>
<td>61 (1)</td>
<td></td>
</tr>
<tr>
<td>Migraine</td>
<td>139 (1)</td>
<td></td>
</tr>
<tr>
<td>Myocardial Infarction History</td>
<td>366 (6)</td>
<td></td>
</tr>
<tr>
<td>Obstructive Pulmonary Disease (Non-Bronchospastic)</td>
<td>973 (16)</td>
<td></td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>180 (3)</td>
<td></td>
</tr>
<tr>
<td>Raynaud’s</td>
<td>21 (&lt;1)</td>
<td></td>
</tr>
<tr>
<td>Sinoatrial Node Dysfunction (no pacemaker)</td>
<td>120 (2)</td>
<td></td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>189 (3)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 and Figure 3 illustrate the relationship between comorbidity number and comorbidity interrelatedness for the entire population that we examined. There was a strong correlation between a patient’s number of hypertension-related comorbidities and his or her comorbidity interrelatedness score ($r = 0.845$). In addition, among patients with two or more hypertension-related comorbidities, there was a strong association between number of comorbidities and number of co-occurring indications and contraindications to the same anti-hypertension class (as occurs when a patient’s diabetes and gout generate both an indication and a contraindication to a thiazide medication). However, as illustrated by the variation in interrelatedness scores across the population in Figure 3, number of comorbidities alone did not fully capture the clinical complexity that was generated by condition interactions. For example, for patients with 4 comorbidities, the interrelatedness score ranged from 4 to 16.
Table 2: Association between Comorbidity Number and Interrelatedness in the Setting of Hypertension Management Decisions (N = 5,997)

<table>
<thead>
<tr>
<th>Number of Hypertension-Related Comorbidities</th>
<th>n</th>
<th>%</th>
<th>Hypertension-Comorbidity Interrelatedness Score Mean (SD)</th>
<th>Number of Anti-Hypertension Classes to Which Patient Has Both an Indication and a Contraindication Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>836</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1,674</td>
<td>28</td>
<td>2.3 (2.0)</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1,584</td>
<td>26</td>
<td>3.9 (2.0)</td>
<td>0.3 (0.5)</td>
</tr>
<tr>
<td>3</td>
<td>976</td>
<td>16</td>
<td>5.0 (1.8)</td>
<td>0.7 (0.6)</td>
</tr>
<tr>
<td>4</td>
<td>492</td>
<td>8</td>
<td>5.9 (1.7)</td>
<td>1.0 (0.6)</td>
</tr>
<tr>
<td>5</td>
<td>257</td>
<td>4</td>
<td>7.0 (1.6)</td>
<td>1.3 (0.6)</td>
</tr>
<tr>
<td>6+</td>
<td>178</td>
<td>3</td>
<td>8.2 (1.6)</td>
<td>1.8 (0.7)</td>
</tr>
</tbody>
</table>

Figure 3. Relationship between number of hypertension-related comorbidities and clinical complexity generated by interactions among conditions (comorbidity interrelatedness)

Discussion

This study demonstrated the feasibility of combining encoded knowledge and EHR data to calculate individual comorbidity interrelatedness scores, and examined the distribution of this construct in a population of patients with elevated blood pressure. We observed a strong relationship between number of chronic conditions and degree of comorbidity interrelatedness. However, variation in comorbidity interrelatedness among patients with similar number of conditions suggests that the construct may capture an important dimension of clinical complexity.

Our work builds on an array of complexity measures that incorporate patients’ chronic conditions into a single index or score. For example, a widely used approach is to reduce multimorbidity to a count of conditions, particularly when researchers desire a simple analytic tool to adjust for variation in comorbidity number. Other measures, such as the Charlson score, Elixhauser, and the Index of Coexisting Disease, rely on a brief list of severe
conditions and have been shown to reliably predict outcomes such as disability and mortality.\textsuperscript{36} However, these measures do not incorporate the additional clinical complexity and risks that may arise when patients’ conditions interact.\textsuperscript{11, 39-41} The comorbidity interrelatedness score that we present here is an attempt to develop an EHR-based measure that incorporates this increased complexity arising from condition interactions.

There are several opportunities to develop the comorbidity interrelatedness construct further to more fully capture clinical complexity. For example, we focused on a single index condition (hypertension) and examined comorbidities that influence decisions about hypertension treatment. An expanded measure could incorporate all of a patient’s medical condition. Algorithms could be applied to weight co-occurring comorbidities that generate an indication and a contraindication to the same drug class, or conditions that are severe and/or identified by providers as having more complicated or time-intensive management requirements. In addition, future measures could incorporate conditions or patient characteristics that do not directly affect specific medication choices but that might influence decisions about management (e.g., physical or mental health conditions, degree of frailty, or social circumstances that influence decisions about number of medications). While these characteristics are not routinely documented in all EHRs at present, there is increasing interest in patient-reported outcomes such as symptoms and functional status that could be incorporated into clinical complexity measures in the future. Developing approaches now with available data will expedite construction of systems that include more variables in the future.

Before such measures, or even the more rudimentary score that we examined in this feasibility study, can influence clinical practice and research, several challenges need to be addressed. Importantly, there is a need to validate the measure to ensure that it captures a unique clinical complexity construct. For example, future studies should examine whether clinical decisions for patients with higher comorbidity interrelatedness scores require greater provider time and effort, particularly when patients have comorbidities that generate concurrent indications and contraindications for the same drug classes. Studies should also explore whether comorbidity interrelatedness is associated with higher rates of clinical inertia (i.e., failure to make treatment decisions due to clinical complexity) or other adverse clinical outcomes. Such studies could be completed using data from an existing CDS tool like ATHENA-HTN, or with an EHR that links patient clinical data to encoded information about comorbidity-drug interactions. Additional work will also be required to consider practical implementation challenges, such as the need to continuously update the KB to ensure that it reflects current evidence about condition-medication interactions and other comorbidity relationships.

Nevertheless, there are a number of potential ways in which a validated comorbidity interrelatedness measure could be used to achieve meaningful use of EHR technology, enhance clinical care, and support research and medical education. For example, understanding the degree to which a patient’s comorbidities affect decisions about medical management could improve case mix adjustment and reward clinicians for providing clinically complex and cognitively demanding care. Some have recommended that performance measures should incorporate patient comorbidities and other characteristics,\textsuperscript{42} and a comorbidity interrelatedness score could be useful for this purpose. In addition, the proposed score and other clinical complexity measures could potentially be used to identify patients who are so complicated that standard disease-focused quality and performance measures should not apply.

**Conclusion**

This study illustrates a method for quantifying clinical decision complexity in terms of comorbidity interrelatedness. This method could be expanded to include additional clinical domains and to include patient data about functional status and social factors, as these data become available in structured format. Incorporating measures of clinical complexity into decision support tools and other data systems could offer opportunities to support clinical care and enhance performance evaluation and research focused on patients with multiple chronic conditions.

**Acknowledgements:** This study was supported in part by Department of Veterans Affairs Health Services Research and Development (HSR&D) grant IMV 04–062 (Collaborative for Improving Hypertension Management with ATHENA-HTN; PI: Goldstein). Dr. Zulman is supported by a HSR&D Career Development Award (CDA 12–173). Dr. Liu’s work on this project was supported by a VA Medical Informatics Fellowship at VA Palo Alto Health Care System. The authors thank Ava Wong for her assistance with literature review and manuscript preparation. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.
References


A SURVEY ON E-PRESCRIPTIONING: AWARENESS, SATISFACTION, BENEFITS AND BARRIERS IN FLORIDA
Pouyan Esmaeil Zadeh PhD, Monica Chiarini Tremblay PhD and Gloria Deckard PhD
Florida International University Miami, FL

Abstract. Electronic prescribing (e-prescribing) allows prescribers to securely send electronic prescriptions directly to pharmacies using specific software [1]. Large national health policy initiatives are promoting the adoption and meaningful use of certified Electronic Health records (EHRs) with electronic prescribing (e-prescribing) capability in order to improve the safety, quality, and efficiency of healthcare delivery. A growing concern among pharmacists is the patient safety implications of new kinds of medication errors and information omissions caused might be introduced by e-prescribing [2]. Currently, pharmacists expect to receive more prescriptions electronically. Thus, they need to be aware of the unintended consequences of e-prescribing. To better understand e-prescribing systems and the potential changes in workflow of pharmacies, it is important to understand what errors are introduced by e-prescribing that did not occur with manual forms of prescribing. This study provides analysis and insight into the awareness of Florida pharmacists and pharmacy personnel of e-prescribing systems, their likelihood of using e-prescribing in their practices, their overall satisfaction with e-prescribing systems, their preference for prescribing, the principal benefits and barriers, and the main errors encountered by pharmacies in Florida.

Methodology. This study uses a cross-sectional internet-based survey administered to pharmacies. The sample consists of pharmacies located in the State of Florida. The study was reviewed and piloted by specialists and consultants before distribution. The survey is distributed online to pharmacies in different settings (such as single site, multiple sites, chain) by the head of the pharmacy association in Florida. The survey is conducted in a two-month window (i.e.: September to October 2014). 197 questionnaires are started, 150 questionnaires are returned completed, the rest are discarded from analysis.

Results. Our findings indicate that most of the pharmacies that responded are moving to adopt and use e-prescribing services in Florida. Almost 91% of our sample states that their pharmacists are equipped with an e-prescribing system and they were using it in their day-to-day practices. A total of 61% respondents indicate that they been using an e-prescribing systems between 1 and 5 years. About 30% state that their pharmacies have been using an e-prescribing system for more than 5 years. Desktop computer turned out to be the most frequently used e-prescribing device for e-prescription (97%), followed by laptop computers (6%), other devices (2%) and handheld devices (1%). A majority (59%) report that their main preference is e-prescription and only 27% indicate that they mainly prefer traditional prescribing. A total of 47% indicated that using e-prescribing system was easy to use and it can be easily incorporated into their practices. Almost 60% of the respondents are satisfied with the current e-prescription system compared to manual forms of prescribing. A total of 90% indicated that they agree with electronic access to complete patient medication histories. These results show the importance of integrated e-prescribing software with an EHR. The majority of the respondents state that the most important benefit of using an e-prescribing system is better and quicker service for patients. Service is improved due to efficiency in pharmacy workflow. The largest barrier to use is the fee paid for each transaction. Errors in e-prescribing are also listed as a barrier. The most common errors observed by pharmacy personnel are: 1) incorrect quantity, 2) wrong dose, strength or frequency, 3) duplicate therapy and unnecessary medical tests, and 4) wrong dosage form. These errors require problem solving by pharmacy personnel. In fact, pharmacy personnel indicate that the e-prescription errors are confusing. They are frustrated and annoyed when the error cannot be easily addressed. The prevalent contributing factors to e-prescription errors are incorrect calculations or data-entry errors. The most common reasons given for the source of the errors is that prescriptions are incorrectly entered by prescribers or that information is incorrectly translated from the prescriber system to the pharmacy system. Finally, the figure shows that four main factors can contribute to concerns about prescribing control substances electronically in Florida. The factors are: factors related to security, privacy and confidentiality of health records, fraud and abuse practice, concerns about time and effort spent for validation and authentication of prescriptions and costs of reporting e-prescribing functionalities, benefits and errors can help pharmacies better take advantage this service. Although e-prescribing can reduce errors related to illegible handwriting the technology itself can cause new types of e-prescription errors. These errors decrease the workflow efficiency because pharmacy personnel have to stop their work and call the prescriber to detect and resolve the errors. Pharmacists and prescribers need to be trained on the function, software interface, and integration of e-prescribing systems into their practice in order to better use of e-prescribing service. Moreover, lack of standardized e-prescribing software is considered as one of the main obstacles affecting the implementation of e-prescribing. Collaborative interaction and communication between prescribers and pharmacies by implementing a standardized e-prescribing system is a possible solution.

Conclusions. Though our study is conducted in the state of Florida, we assert it is generalizable to other US states since Florida is both geographically and ethnically diverse. However, we plan to expand our study to other states. This study is important because building awareness of

References.
Design, Development, and Initial Application of a Systematic, Semi-Automated Predictive Analytics Framework for Health Care

Samir E. Abdelrahman, PhD, Kensaku Kawamoto, MD, PhD, MHS
Department of Biomedical Informatics, University of Utah, Salt Lake City, USA

Introduction

Predictive analytics can be leveraged in health care in various ways to identify risk factors, predict outcomes, and manage patients in an individually tailored, risk-stratified manner. For example, predictive analytics can be used to identify patients hospitalized with congestive heart failure who are at high risk of readmission, so that these high-risk patients can be monitored closely post-discharge. Despite its significant potential, however, predictive analytics is not widely leveraged in most health care settings. One of the main barriers to the widespread application of predictive analytics is the amount of expertise and time required. To address this important challenge, University of Utah Health Care sought to develop a systematic predictive analytics framework for health care, with steps in the process automated wherever possible. The current state and application of this framework are presented here.

Methods

Clinical and operational leaders at University of Utah Health Care were consulted to identify priority use cases for health care predictive analytics. Several pilot projects were completed in these areas, with each project leading to the development and refinement of the predictive analytics framework based on the use case requirements. Processes amenable to automation were identified and implemented on an iterative basis.

Results

Figure 1 outlines the predictive analytics framework, which includes a model building stage (top part of Figure) and an operational deployment stage (bottom part of Figure). Both structured and unstructured electronic health record (EHR) data may be used, with a present focus on structured data. Model building entails data collection and integration, data analysis, systematic model development, and risk factor analysis, while operational deployment entails implementation of the back-end infrastructure as well as the user-facing intervention. Clinical experts are essential for several of the steps in the framework, including the identification of potentially relevant risk factors and the design of clinical interventions. Automation has been incorporated primarily into the model building aspect of the framework, so that this aspect of the framework can largely function as an automated pipeline.

The framework has been applied to several use cases, including the prediction of congestive heart failure readmission, the prediction of clinical appointment no-shows and last-minute cancellations, the prediction of complications from medical procedures, and the prediction of patient satisfaction. Model performance has generally been strong. For example, this framework was able to predict congestive heart failure readmissions with a positive predictive value of 50% and a negative predictive value of 96%, with the area under the curve significantly higher than in previously published models.

Conclusion

A systematic and largely automated predictive analytics framework can be successfully implemented in support of efforts to improve health and health care.

References

Building of Community Health care Data Bank (EHR) using positional and temporal tracking and collecting DATA

**Keywords.** Objective Data Base, EHR, Spatiotemporal tag

Masakazu AIHARA1, Sizukiyo Ishikawa1, Masanobu OKAYAMA2, Yasutomi KINOSADA3, Masayuki HONDA4, Yoiti SEKI5, Yoshikazu NAKAMURA1, Eiji KAJII1
1 Center for Community Medicine, Jichi Medical University
2 Kobe University Graduate School of Medicine
3 Gifu University Graduate School of Medicine
4 Division of Medical Informatics, Nagasaki University Hospital
5 Gunma University Graduate School

**Introduction**
The data concerning community health, such as, medical health records, health check-up, cancer screening, nurse care services have been separately stored in each their own database. The difficulty of sharing information between these regions, was in the compatibility of data. The aim of this study was to develop the Community Health Databank (EHR) for providing an efficient community health care.

**Methods**
We developed a novel electronic system utilizing a spatiotemporal tag. Used by adding a space-time tags, without relying on the existing ID, all data pertaining to an individual, it may be data aggregation on the timeline. (Figure1)
Furthermore, the developed databases is not have personal information, has a structure that can be used with confidence. (Figure2)
Using this system, we obtained PHR (personal health records) including personal individual behavioral records, medical records, and health records.
We assessed whether these records integrated appropriately and safely with independence of private information, such as name, date of birth, and address. (Figure3)

**Results**
Our system succeeded in storing the various individual health records comprehensively. Using tagged data, any information regarding health care can be shared with protection of the private information regardless of their format and style.

**Discussion**
The Community Healthcare Databank would enable an integrated approach for health care information without treating private information. Then, utilization of this bank could help to establish a standardized health care system.
Knowledge Base Acquisition For Rare Concepts Using Manual Bootstrapping
Patrick R. Alba, BA1,2, Scott L. DuVall, PhD1,2, Joanne LaFleur PhD1,2, Adam P. Bress, PharmD, MS2, Olga V. Patterson, PhD1,2
1VA Salt Lake City Health Care System; 2 University of Utah, Salt Lake City, UT

Abstract
Concept identification is an essential task of any natural language processing (NLP) system built to perform information extraction from clinical documents. It typically relies on term identification and context disambiguation. The lack of a comprehensive lexicon necessitates corpus-based knowledge base acquisition. A major challenge arises in cases when terms that represent the concept of interest are ambiguous and the concept is rarely mentioned in text, which makes manual annotation not feasible and automatic bootstrapping less efficient. We propose a method of computer assisted knowledge base acquisition that requires less effort than a manual review, and achieves a more accurate outcome than an automatic approach would in these cases.

Background
The Canadian Cardiovascular Society (CCS) angina grading scale is a generally accepted method to describe a patient’s angina severity1. An NLP system aiming to extract CCS scores from clinical notes requires an accurate and comprehensive lexicon and disambiguation model that reflect the target corpus. While the CCS scoring system is commonly used to describe angina, explicit mentions of specific scores are rare. The issue of identifying mentions of the score in documents is complicated by the highly ambiguous terms used by clinicians. For example, the abbreviation CCS is more frequently used to describe cubic centimeters rather than the Canadian scoring system. (e.g. 1000ccs of water). The two knowledge base acquisition approaches that are typically used are manual annotation or semantic bootstrapping. However, annotating a large set of random documents would not be feasible due to the rarity of the relevant concept. Automatic bootstrapping requires highly precise seed words, thus, not appropriate in this case due to the ambiguous nature of the known terms2.

Methods
The original CCS guideline provides only two relevant terms associated with the concept of interest, Canadian Cardiovascular Society score and CCS score. All available documents for a random set of 10,000 patients were retrieved from a large national database. All instances of the seed words were identified using keyword search and manually reviewed in order to discover co-occurrence patterns used by clinicians. The patterns were then iteratively applied to the full set of documents without the seed words. As new terms were discovered, they were added to the seed word list. The patterns collected through this manual bootstrapping approach were used to create or modify a series of regular expressions. Several iterations of the process were completed before no new co-occurrence patterns were discovered. These regular expressions reflected one of two types of context patterns: those that reflect the concept of interest, and those that indicate that the term represented a different concept, thus serving as a disambiguation patterns. The two sets of regular expressions served as the knowledge base for the CSS score extraction system.

Results
Using this approach, all instances of a CCS score within 1.8 million documents were reviewed by a single person in under 40 hours. While the seed words were very common, this discovery process identified that only 0.02% of the documents contained the concept of interest. The described approach was able to uncover not only the rare cases of mentioned concepts but also instances where the concept was implied. For example, the term “CCS” is implied in instances like “angina class 1” and “stage IV angina”. Though largely manual, this corpus-driven approach resulted in a comprehensive knowledge base for information extraction with limited expert input.

Conclusion
The amount of person-hours required in order to build a knowledge base using the described manual bootstrapping method is considerably less than other standard manual methods and more precise than computational methods.

References

Acknowledgement
This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204 and the University of Utah.
Personalized medicine beyond genetics: using personalized model-based forecasting to help type 2 diabetics understand and predict their post-meal glucose

David Albers¹, PhD, Matthew Levine¹, Bruce Gluckman², PhD, George Hripcsak¹, MD MS, and Lena Mamykina¹, PhD

¹Department of Biomedical Informatics, Columbia University, New York, NY
²Department of Engineering, Pennsylvania State University, College Park, PA

Introduction
Type 2 diabetes (DM2) is a disease with dire consequences if not treated to minimize its effects, is costly ($245 billion a year in 2012), and is prevalent affecting over 8% of the US population. Treatment of DM2 relies primarily on nutrition, exercise, and medication. Because of this, self-management is an essential part of DM2 care. However, DM2, as a disease, is highly dependent on the individual and while there exist general guidelines for healthy eating, each individual with diabetes must find their own way to translate these guidelines into daily meal choices. This is usually accomplished through a lengthy, often complex and opaque, trial and error process in which individuals experiment with different foods and observe their impact on blood glucose levels. Moreover, as is the case with many complex nonlinear processes ranging from aircraft flight to the management of large industrial machinery to weather forecasting, humans are not well suited to finding context-specific (i.e., personalized) optimal solutions without support. The methodology we present here, model-based forecasting (MBF), a borrowed technology broadly termed data assimilation, is the machinery that makes it possible for personalizing DM2 treatment forging a personalized, glucose/insulin/nutrition based forecast of post-meal, continuous glucose forecast. This methodology through the to personalization of mechanistic models that allow for nutrition-based post-meal glucose forecasts, creates a concrete step between a nutrition choice and variables correlated with long-term health, high glucose and HA1C. While difficulties remain, the data assimilation methodology will form the computational foundation for personalized application-based DM2 interventions such as mobile applications that DM2 patients use for making meal choices and for managing and observing their overall health.

Methods
We prospectively collected 21 days of nutrition, glucose, and HA1C data from patients in a DM2 intervention study. We then selected two mechanistic endocrine models, a relatively simple an ultradian model developed to model fast glucose-insulin dynamics in healthy individuals and a complex meal model developed as a detailed model of meal-time metabolism as related to glucose-insulin dynamics for both individuals who were either healthy or had DM2. Both models can take glucose, insulin, and nutrition as input variables and can forecast glucose, insulin, and between 30 and 70 other physiologic parameters. To integrate an individual’s data, personalize the models through parameter training, and generate a forecast, we embedded the models in a dual unscented Kalman filter (UKF), a DA methodology used to optimize and forecast nonlinear processes. Using the patient data we then created glucose and nutrition based post-meal glucose and insulin forecasts for both models. To evaluate which model performed better we use both pointwise predictive density of the model forecast compared to measurement and KL divergence between the distribution of glucose measurements and forecasts.

Results
After the initial training of 20-30 meals, both models personalized via convergence to the individual patient and both models forecast post-meal glucose well. The ultradian model minimized the point-wise predictive error and therefore provided a better immediate post-meal forecast. In contrast, the meal model minimized the KL divergence between kernel density estimate of the measured glucose and forecast glucose densities, implying that the meal model provided a better representation of glucose representation of the patient in general, while not providing as optimal of an immediate post-meal glucose forecast.

Discussion
The MBF-based methods, even with limited training, can provide accurate, personalized, nutrition-based glucose. The hope is that MBF methodology will be used to personalize nutrition-based, post-meal glucose forecasts, allowing individuals with DM2 to make personalized, informed nutrition choices based on immediate, concrete meal-dependent glucose forecasts combined with an understanding of how glucose levels and dynamics affect health. Nevertheless, serious questions remain. While both models personalized to the individual, they provide different “optimal” knowledge. The meal model yielded a better overall picture of glucose dynamics while the ultradian model provided a better immediate post-meal glucose forecasts. We do not know which type of knowledge would be more useful for patients to make meal choices and manage their DM2. Given the 10’s of mechanistic glucose models available we must evaluate which model will be of greatest potential impact. But, to choose an evaluation metric we must understand what information about future glucose dynamics, when provided to a patient, will yield the best self-care-based outcome.

Acknowledgments
We acknowledge NLM grant R01 LM06910 and NIDDK grant 5R01DK090372
Increase in Prescriber Error Rates Following Implementation of Computerized Physician Order Entry

Gretchen B. Alexander, M.D., Gilbert M. Ramos, M.A., Maricopa Integrated Health System, Phoenix, AZ

Background
A critical impetus for the adoption of computerized physician order entry (CPOE) systems has been the anticipated improvement in patient safety through reduction in medication errors. A number of studies have demonstrated an association between CPOE adoption and decreased medication errors [1]. However, improved outcomes with respect to patient mortality and morbidity have not been as clearly described. Other work has shown unexpected consequences of CPOE including increases in physician cognitive workload [2] as well as increases in new types of medication errors [3]. It has been hypothesized that CPOE with clinical decision support (CDS) would reduce prescriber errors, but few studies have specifically examined the effect of CPOE adoption on prescriber error rates.

Purposes
To evaluate the effects of a commercial computerized physician order entry (CPOE) system on the rate of medication errors in a public hospital, and to compare errors by type and degree of patient harm prior to and following CPOE implementation.

Methods
A retrospective longitudinal review of data from the hospital’s Quality and Risk database (MIDAS) was performed. Error rates were collected and evaluated for a 36 month timeframe; 14 months before CPOE implementation (pre) and 22 months following CPOE implementation (post). Percentages were calculated for all error types (source of error) and categories (degree of associated harm). Comparisons were performed with Student’s T or Chi squared analysis, as appropriate. Statistical significance was defined as p<0.05 for all calculations.

Results
A total of 2644 errors were self-reported during the study timeframe. Overall, self-reported error rates did not change statistically following CPOE implementation (1295 pre and 1349 post). However, individual error classes did show statistical change with increases in some error types and decreases in others. Significant increases in prescriber errors and errors associated with a higher level of patient harm (Type D and E errors) were observed.

Errors in the following categories showed a statistically significant increase pre to post:
Category A (8.3% pre and 31.6% post; p<0.0001)
Category D + E errors (3.32% pre and 5.71% post; p<0.05)
Administering (43% pre and 47.7% post; p=0.03)
Prescriber error (5.3% pre and 15.3% post; p<0.0001)

Conclusions
While error rates may also be affected by other factors, computerized physician order entry adoption in a public hospital appeared to be associated with increases in certain types of medication errors. The unintended consequence of increased prescriber errors suggests continued need for research on best practices for: consideration of clinician workflow during system development; physician training and support during and after CPOE implementation, and physician involvement in ongoing system optimization.

References
Identifying Home Care Clinical Practices Most Associated with Hospital Readmissions and Non-Admitted ER Visit Rates: Secondary Data Analysis

Dari A. AlHuwail, MSc, Güneş A. Koru, PhD
1University of Maryland, Baltimore County, Baltimore, Maryland, USA

Introduction
Utilization outcomes, particularly, the (i) hospital readmission rates and (ii) non-admitted ER visit rates have become a subject of growing interest in recent years due to cost concerns [1][2]. Arguably, as the health-care processes of home health agencies (HHAs) improve, their monitored utilization outcomes should improve as well [3]. However, due to limited resources, many HHAs need to apply focused and prioritized process improvement initiatives [4]. To support such initiatives, this study identified the home care practices most associated with higher utilization rates.

Methods
This research followed a secondary data analytics approach. Sixteen process variables about adherence with clinical practices and two utilization outcome measures, all represented as rates, were obtained from 2014 Medicare Home Health Compare (MHHC) Database. The data set included approximately 10,000 observations, each corresponding to a single HHA. To control for socioeconomic status and ruralness, two surrogate measures, Median Household Income and Rural-Urban Commuting Area (RUCA) were attached to each observation based on the HHA’s zipcode. RUCA was coded as an ordinal variable, which increased with ruralness. RUCA and Median Income were obtained from University of Washington and University of Michigan, respectively. Regression trees were developed using a computationally intensive process facilitated by the “rpart” package in the statistical environment, R. Two types of regression trees were developed by taking (i) hospital readmission rate and the (ii) non-admitted ER visit rate as the response variables. For both types of models, the predictors consisted of the process variables, median income, and RUCA.

Results
Interestingly, RUCA explained the utilization variance most in all models. Both utilization outcomes were worse for HHAs serving in rural areas. This finding shows that including RUCA as a control was appropriate. Median Income became relevant only when RUCA was not included. For hospital readmissions, the following three clinical practices, listed in the decreasing order of explaining variance, were most important: (1) How often the home health team checked patients’ risk of falling, (2) How often the home health team began their patients’ care in a timely manner, and (3) With diabetes, how often the home health team got doctor’s orders, gave foot care, and taught patients about foot care. For non-admitted ER visits, a similarly ordered list of top-three clinical practices was (a) checking for fall risks, (b) treating patients for pain, and (c) starting care in a timely manner.

Informatics Perspective
Investments made for Home Health Compare database seem to be useful for supporting reproducible research. While advanced computational tools are readily available today, collection and integration of quality data from multiple sources at the desired level of granularity seem to be the major challenges in the near future. From an organizational process improvement perspective, such integration through the development of Health IT tools can inform HHAs to improve clinical quality, safety, and patient outcomes.

Conclusions
A smaller number of clinical home care practices were most associated with two important utilization outcomes. Checking for fall risks and timeliness, also important practices for patient safety were associated with both hospital readmissions and ER visits. These results should be immediately useful to HHAs in their focused and prioritized process improvement efforts.

References
**DECISION ANALYSIS FOR OROPHARYNGEAL CANCER IN RADIOTHERAPY**

Ahmad K. Aljadaan¹, John H. Gennari¹, Mark H. Phillips², and Wade P. Smith²

¹Department of Biomedical Informatics and Medical Education; ²Department of Radiation Oncology, University of Washington, Seattle, WA

**Abstract:**
In radiation oncology, for some populations of patients, there is uncertainty about cure rates for a given level of radiation. Because the dose-response curve is not precisely understood, one cannot predict exactly how different prescription doses might affect outcomes. Our goal is to understand where the decision points are: When does more information change clinical practice?

**Background:** In radiation oncology, there is sometimes insufficient data relating treatment dose levels to patient utility. The National Comprehensive Cancer Network provides clinical practice guidelines, but the optimal dose level (number of greys of radiation) is still sometimes unclear, especially for specific sub-populations of patients. Ideally, different dose levels would be associated with different cure rates, but the shape of this dose-response curve is not fully known. Our goal is to apply decision analysis to support both decision making and potentially the design of future clinical trials, so that better, more evidence-based decisions can be made about prescription dose levels for specific patients.

In Oropharyngeal cancer, side effects such as xerostomia (dryness of mouth) and dysphagia (difficulty swallowing) often occur as a result of the radiotherapy treatment with the standard dose level. These side effects are severe and can induce life-long social and emotional burdens. A larger dose can increase the probability of disease control but also increase the probability of side effects. Recently, studies found that Human Papillomavirus positive (HPV+) populations of oropharyngeal cancer patients have a higher survival rate after therapy. Current randomized controlled trials are exploring the reduction of radiation dose to minimize side effects for HPV+ patients. Treatment de-escalation for these patients is also being explored through a less-toxic systemic therapy (Cetuximab) as an alternative to standard chemotherapy.

**Methods:** We built a decision model incorporating an Influence Diagram and a Markov Model for outcomes for reduced treatment in oropharyngeal cancer. The Influence Diagram calculates probabilities of outcomes which are used in a Markov cohort simulation in order to calculate Quality-adjusted Life Years (QALYs) of patients undergoing radiation therapy and systemic therapy. In order to change clinical practice the QALYs for reduced therapy must be higher than for standard therapy. We also performed sensitivity analysis to determine the robustness of the model for certain parameters. In Fig. 1, we show four different curves representing different assumptions about the rate of cure per increasing radiation dose (as % change per Gray).

**Impact:** Our goal is to delineate which outcomes of clinical trials would lead to changes in clinical practice, and to use Value of Information analysis to identify which model parameters are most crucial to decision making. This type of model has broad applicability to other disease sites in radiotherapy and to clinical trials in medicine in general.

**Figure 1.** Preliminary Results for Expected Utility after Radiotherapy for HPV+ late stage (III/IV) Oropharyngeal Cancer. The Utility decreases as the disease control rates fall, and increases as toxicity rates fall.
Medications and Events Most Commonly Discussed in Facebook and Twitter

Heidi G. Bell RPh, MD¹, Lorrie Schifano, PharmD², Laurie S. Anderson, PharmD², Gregory E. Powell PharmD, MBA², Hoa Van Le, MD, PhD², Harold Rodriguez³
¹ZeroChaos, RTP, NC; ²GlaxoSmithKline (GSK), RTP, NC; ³Epidemico, Boston, MA

Introduction
As social media rises in popularity and in use as a communications tool, pharmaceutical industry and regulatory agencies alike are working to see how social listening (the process of identifying and assessing what is being said about a company, individual, product or brand in social media) can be used to augment the current data sources for pharmacovigilance (active monitoring of drug safety). In working towards better patient safety regarding drugs and other pharmaceutical products such as vaccines, we set out to determine what information is available through social listening that might help determine best practices for use of these data.

Methods
All publicly available English language posts in Facebook and Twitter from November 2012 to February 2015 which mentioned a pharmaceutical product (medication, vaccine, or consumer healthcare product) and at least one symptom that resembled an adverse event as tagged by an automated classifier system (“Proto-AE”) were obtained and de-identified by third-party vendor Epidemico™. Post content was acquired through public application programming interfaces (APIs, the Twitter Public API, e.g.) and data provisioning services (e.g. DataSift). Post text was tokenized and tokens identified by Natural Language Processing (NLP) methods. Based on identified tokens, products and symptoms were associated with each post. Each product name was mapped to standard Anatomical Therapeutic Chemical (ATC) classifications, and each Proto-AE was mapped from vernacular language to MedDRA terminology using a custom dictionary developed for Epidemico’s MedWatcher Social™ platform. The products and symptoms were then counted and summarized by export into a spreadsheet.

Results
A total of 862 products were mentioned in the context of a possible adverse event over the time period of collection, with 1,411,705 total Proto-AEs among them. Many posts paired more than one symptom with a product, allowing for 6,796,179 total MedDRA PT symptoms coded among 751 individual PTs.

The most common products by ATC Level 1 (listed along with most common products within the class) were: Nervous system (narcotics and acetaminophen, ADHD, psychiatric, antiepileptics), 749,292 Proto-AEs (53%); Respiratory system (antihistamines, inhalers, antitussives including hydrocodone/codeine), 245,485 (17%); Antiinfectives for systemic use (vaccines, anti-influenza, antibiotics), 204,556 (14%); Musculoskeletal system (NSAIDs, botox, muscle relaxants), 52,822 (4%); and systemic hormonal preparations (corticosteroids, thyroid hormones) 35,448 (3%).

The most common Proto-AE symptoms by System Organ Class (SOC) in the MedDRA hierarchy (listed along with top three PTs in class) were: General disorders and administration site conditions (pain, malaise, drug ineffective), 2,070,364 (30%); Nervous system disorders (headache, altered state of consciousness, memory impairment), 1,169,018 (17%); Psychiatric disorders (depression, anxiety, insomnia), 1,150,096 (17%); Respiratory, thoracic, and mediastinal disorders (asthma, dyspnea, oropharyngeal pain), 333,702 (5%); and Injury, poisoning, and procedural complications (overdose, injury, incorrect dose), 315,361 (5%).

Conclusion
Social listening provides a large number of Proto-AE reports even when limited to just these two popular social media sites and the English language. In looking at the use of the social listening and considering how it may best augment current pharmacovigilance practices, it is important to consider that the top five therapeutic classes and Proto-AE symptoms account for 91% and 74% of overall posts, respectively. Overall, social listening offers the potential to supplement existing pharmacovigilance practices. More research is needed to understand the broad applicability of this emerging capability.
Enhancing Use of the Problem List in the Inpatient Setting

Charles H Andrus MHA 1, Kevin O’Bryan MD2, Phillip Asaro MD 2, Stanley Paul Hmiel MD PhD2, Feliciano Yu MD MSPH 2
1St. Louis Children’s Hospital, St. Louis MO, and 2Washington University School of Medicine, St. Louis MO

Introduction

Problem lists management modules have become standard in electronic health records (EHRs). Despite this, many clinicians find these modules to be of limited utility because of inability of the system and ontology used to meet the needs of clinicians. Problems lists need to provide utility through being accurate, complete, and non-redundant to meet the needs of clinicians.1

St. Louis Children’s Hospital has a best of breed environment where our inpatient care is primarily performed using a separated EHR from outpatient care. The EHR vendor provided problem list design was initially rolled out with minimal modification but usage was extremely low and as electronic documentation increased, the need to enhance problem list management grew. An approach was undertaken to enhance the user interface and functionality of the problem list management module. International Medical Objects (IMO) ontology was selected as the ontology for the problem list and heavy clinician engagement was used to redesign a user interface.

A Novel User Interface

The user interface was redesigned to employ a graphic user interface with the ability to drag and drop problems and have “quick buttons” to perform common actions such as resolving a problem or evolving a problem from a generic problem such as fever to a specific problem later in the clinical course such as pneumonia. In addition, we wanted to ensure that accurate problem lists were sent out at the end of visits to provide continuity of care on an outpatient basis and to ensure the patient’s chart would be as accurate as possible in preparation for a future visit. This was accomplished by a similar user interface where users would create discharge diagnoses and chronic problems from the current problem list. Users are forced to address all acute issues as either resolved, discharge diagnosis, and/or chronic diagnosis. Once submitting this, users are presented with a preview of what the patient’s problem list will look like at the next visit and the ability to edit this as needed to ensure that future visit information is accurate.

Adding Problems through Search Tools

The second major change we made was to enhance the ability of the system to support users looking for specific problems. IMO’s ontology is designed to contain a complete list of problems in clinician friendly terms with synonyms for any terminology a clinician might be likely to use. While this is successful when clinicians search for very specific terms, generic terms produce long lists of results that are difficult for a clinician to review and select an appropriate term. As a result clinicians select relatively generic terms. Initially we managed this by providing an alternative list of the top 15 problems other clinicians had selected while using similar search terms. This improved user satisfaction, but did not lead to the use of more specific terms; rather it augmented poor term selection.

To resolve this issue, we developed a new search algorithm. In this algorithm, the computer counts frequency of each word in all of the search results. Terms present in >95% of results or <5% of results are excluded. Certain common words are also eliminated. The computer then presents these terms in order of decreasing frequency to the user as possible filters for the user to use. The user can use these terms as include or exclude filters with one click. Using this functionality clinicians rapidly narrow from hundreds of terms to typically 2-4 terms with 2-3 clicks. Additionally, users learn terminology for future searches and include these terms in their initial search in the future.

Conclusion

Since implementation of these solutions, we have seen enhanced usage of the problem list and user acceptance. The average visit now has 6 (±6.13) problems, rather than the previous 3 (±3.03).

References

Feasibility and Acceptability of an Online Maternity Education Platform

Adriana Arcia, PhD, RN1, Deren Bader, CPM, DrPH2, Eva Rose Asaan Warner, BA1
1Columbia University, New York, NY; 2Maternity Neighborhood, Charlottesville, VA

Background and Significance
There exist stark racial, ethnic, and socioeconomic disparities in US childbirth outcomes.1 Uptake of prenatal care, abstinence from substance use, adequate intake of folic acid, and weight gain within recommended guidelines are among the behaviors promoted by Healthy People 2020 because they are known to improve birth outcomes.2 Women at greatest risk for poor outcomes because of social factors also are more likely to have low health literacy and thus be disadvantaged in their ability to receive and interpret health messages that promote healthy self-management.3 However, health literacy is not solely a function of individuals. Health literate organizations “include populations served in the design, implementation, and evaluation of health information and services” and “meet the needs of populations with a range of health literacy skills while avoiding stigmatization.”4 Therefore, a health literate maternity information system prioritizes identifying and meeting the unique needs of the women most at risk for poor outcomes.

The Maternity Neighborhood™ (MN) Care Guide™ is a mobile-ready, web-based patient education platform populated with curated, evidenced-based, multi-media content for maternity care. Content is automatically delivered week-by-week based on gestational age. Secure messaging is an integral part of the platform in that it enables communication about content between the pregnant woman and her care team, thus promoting engagement with care. The goal of the research in progress described here is to insure that Care Guide™ (known as Maternity Information Access Point or MIAP for research purposes) is a health literate resource for at-risk women. To that end, the aim of this study is to explore the feasibility and acceptability of an online maternity education platform among Medicaid-enrolled women. The theoretical framework is Coiera’s common ground theory of healthcare communication which posits that “for computational tools to be of value, they have to share ground with human beings. Users need to know how to use the system and the system needs to be fashioned to users’ needs.”

Focus Group & Field-Testing Protocol: We will recruit 24-32 pregnant (aged 18+, ≤35 weeks’ gestation) Medicaid-enrolled women via community-based doula groups to engage in pairs of focus groups bookending a one-month at-home MIAP field-testing period. Initial group topics are: 1) barriers to/facilitators of Internet access; 2) Internet use, including social networking, for maternity information; 3) preferred forms of maternity information access; 4) online platform acceptability; and 5) desired attributes of platform content. During field-testing, participants will be encouraged to meet information needs with MIAP and engage with the research team through secure messaging. In follow-up groups, usability and satisfaction will be evaluated qualitatively and with items modified from the Unified Theory of Acceptance and Use of Technology (UTAUT2). Content analysis will be used to analyze qualitative data.

Discussion
Findings will be used to identify opportunities for facilitating Internet access for Medicaid-enrolled women within a prenatal services context (e.g., subsidies for data plans) and to configure a health literate platform that meets the needs of the target population. Planned future work will evaluate the impact of platform usage on clinical outcomes.

Acknowledgments
Columbia University Provost’s Grants Program for Junior Faculty Who Contribute to the Diversity Goals of the University

References
A Smartphone-based Gait Assessment System for the Elderly

Naoya. Arisaka, Msc¹, Kosuke. Mizuno, PhD², Noritaka. Mamorita, PhD¹, Yoshitaka. Shiba, PhD³, Harukazu. Tsuruta, PhD¹

¹Department of Medical Informatics, Kitasato University, Sagamihara, Japan
²Kitasato University East Hospital, Sagamihara, Japan
³Department of Physical Therapy, Kitasato University, Sagamihara, Japan

Introduction

The kinematic assessment of posture and movement in walking plays on important role in rehabilitation for the elderly and in the assessment of prosthetic replacement arthroplasty. However, the quantitative evaluation of the posture and movement usually requires expensive and large-scale apparatus such as the 3-dimensional position analyzer and therefore is limited to laboratory settings. To solve this problem, we developed a smartphone-based system which measures and analyzes the motion in walking using the built-in sensors and communication via Bluetooth Low Energy.

Methods

The system is composed of measuring devices and a control device. The application is developed using Objective-C and we used iPod touch (Apple Inc., USA) in this study. To verify the accuracy of our system, optical data were gathered simultaneously using an optoelectronic motion capture system. Using the developed motion measurement system, we measured and evaluated the posture of sixty elderly people and fifty patients with Parkinson's disease periodically. The tilt angles of the trunk and pelvis were measured by two iPod touches which were synchronized with each other. This study was approved by the Ethics Committee of Kitasato University Hospital and all participants provided their informed consent to participation.

Results

The control device can order start/stop of the measurement to measuring devices simultaneously (Figure 1) and both devices can display the summarized results in addition to the observed changes of the three-axis posture angle and acceleration in a graphical manner (Figure 2). The observed data were automatically uploaded to the server. We compared the posture angle in walking between the older (71.1 ± 5.1 years, old (mean±SD), n = 44) and the younger (21.3 ± 1.0, n = 30) women. The trunk of older women was significantly tilted forward, pelvis was significantly tilted backward, and the left-right movement of the pelvis was greater than the younger women.

Conclusion

We developed a smart-phone based motion measurement and analysis system. The results suggested that we can use our system instead of the expensive optoelectronic motion capture system. The difference of observed posture angle between our system and the optoelectronic system was less than 2.3 degree¹. We are now collecting data of 110 subjects including patients with disabilities at four hospitals to obtain reference values for the rehabilitation of the elderly. The measuring device will be used as a stand-alone measuring tool that the elderly can use at home.

References

Analysis of Computerized Clinical Reminder Activity and Usability Issues

Shazia Ashfaq, MD, MBA1,7, Kristin M. Bell, MD1,2, Megan Difley1, Sara Mortensen1, Kellie Avery1, Steven Rick, BS1,7, Nadir Weibel, PhD1,8, Braj Pandey, MD1,2, Charlene Weir, PhD9, Harry S. Hochheiser, PhD9, Yunan Chen, PhD9, Jing Zhang, MS12, Kai Zheng, PhD11, Richard Street, PhD5,13,14, Mark T. Gabuzda, MD1,2, Neil Farber, MD3, Lin Liu, PhD1,3, Alan Calvitti, PhD 1, Zia Agha, MD, M.S4,7,12

1VA San Diego Healthcare System, San Diego; 2Departments of Internal Medicine and 3Family and Preventive Medicine, UC San Diego; 4West Health Institute, San Diego, CA; 5Department of Communication, Texas A&M University; 6Department of Medical Informatics UC Irvine; 7Veterans Medical Research Foundation, San Diego; 8Department of Computer Science and Engineering, UC San Diego, CA; 9Department of Biomedical Informatics, University of Pittsburgh; 10Department of Biomedical Informatics, University of Utah; 11University of Michigan, Ann Arbor, Michigan; 12University of California San Diego, San Diego, CA; 13Baylor College of Medicine, Houston, TX; 14Houston VA Center for Innovations in Quality, Effectiveness and Safety

Introduction

Computerized Clinical Reminders (CCRs) play an integral role within the Computerized Patient Record System (CPRS), the Veterans Health Administration’s Electronic Medical Record (EMR). CCRs encompass clinical documentation and administrative tasks for improving quality of care and safety while adhering to institutional guidelines[1]. Our study aims to objectively assess the burden of CCR work during clinic visits at the VA San Diego.

Materials and Methods

We recorded clinical workflow using an eye tracker, a video camera, and usability software installed on the physician’s computer to record the physician’s gaze, their computer screen, and mouse activity. We subsequently tagged click sequences manually in order to connect them to reminder activity. Qualitative data about CPRS use and CCRs was collected through interviews. We grouped the named CCRs and rank ordered them by frequency and time-at-task. We selected 2 reminders (Tobacco cessation and Medication reconciliation) for task analysis. Video review and content analysis was performed manually. Physician gaze activity was analyzed and displayed graphically as gaze paths and heat maps using proprietary eye tracking software.

Results

We analyzed CCR activity data for n=89 outpatient visits (Fig) with 16 physicians (9 primary care and 7 specialists). 148 reminders completions, (Median=0, IQR=[0-3]), all PCPs, were concentrated in 37 visits (Median=4, IQR=[2,6]). 21% physicians reported they work on reminders after the visit (post work). We also analyzed CCR activity for new versus established patients. We found 29 distinct types of CCRs and the most frequent 3 were medication reconciliation, HIV screening, and colorectal cancer screening. The average time to complete a reminder was under 1 minute. Qualitatively, physician interviews indicate burdensome and unnatural CCR workflow is associated with: perceived time pressure, redundant documentation (“see notes” in reminder text box and auto-appended reminder template text in notes), and blocking states in the EMR (inability to search for information with reminders open).

Conclusion

Based on observations of CCR activity, time-at-task is small relative to the length of visit. However, in other metrics, CCR burden is reflected by physician’s perception of time pressure attributed to CCR work, which appears to arise from unoptimized clinical documentation system and CCRs. Redesign of reminder functionality and revision of content could improve the usability of CCRs for physicians.

Acknowledgements

Funding by 5RO1 HS021290-03n (AHRQ): Quantifying EMR Usability to Improve Clinical Workflow.

References

Evaluation of a Self-Triage Decision Aid System in Pregnancy-Induced Hypertension and Diabetes Mellitus: Preliminary Results of a Randomized Control Trial

Azam Aslani, PhD1, Fatemeh Tara, MD2, Leila Galichi, PhD3, Sina Madani, MD, PhD4, Ameen Abu-Hanna, PhD5, Saeid Eslami, PhD1,5

1Department of Medical Informatics, Mashhad University of Medical Sciences, Mashhad, Iran; 1Department of gynecology, Mashhad University of Medical Sciences, Mashhad, Iran; 2Mental Health Research Center, Iran University of Medical Sciences, Tehran, Iran; 3The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX, USA; 4Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands

Abstract
It has been shown that access to health information can benefit patients by helping them making a more knowledgeable decision in the clinical care process. Computerized decision aids can be used to facilitate such decisions. We investigated the effect of a self-triage decision aid tool on enabling pregnant women to monitor the development of gestational hypertension and diabetes mellitus and actively participate in their health care process.

Introduction
Acquisition of reliable and actionable health knowledge can help patients actively participate in their care process and make more effective decisions about medical interventions. Published reports in Iran showed that two important causes of maternal death are the lack of awareness of alarming symptoms during pregnancy and ignoring recommendations from the health care providers. Pregnancy-induced hypertension (PIH) is the cause of death in 17% of reported cases. It has also been shown that gestational diabetes mellitus (GDM) can cause macrosomia and very low blood glucose levels in newborns. We anticipated that pregnant women can quickly identify the alarming symptoms and prevent high-risk situations by using a self-triage computerized decision aid (CDA) tool. The objective of this study was to evaluate the effect of a self-triage decision aid system on pregnant women in managing PIH and GDM during pregnancy.

Methods
Randomly selected pregnant women between the ages of 16 and 38 were recruited in a randomized control study. They were examined by six gynecologists in two publicly funded maternity health service centres and by two other gynecologists in private offices between December 2013 and June 2014. The unit of randomization was the gynecologist. Both the intervention and the control groups received the conventional care while CDA was used only by the intervention group. We mainly measured self-efficacy by self-efficacy scale, knowledge by 15 multiple choice questionnaire and anxiety by Spielberg state anxiety inventory. The authors who designed the questionnaires assessed CDA utility, ease and frequency of use.

Results
All 116 participating pregnant women (56 in intervention group and 60 in control group) completed their pre and post pregnancy evaluations. There was no significant difference between the two groups in terms of their age, gestational age, number of previous pregnancies and birth, and planned pregnancy. Improvement in the reported feel of self-efficacy in the intervention group for PIH and GDM were 15% and 11.2% respectively (both with p=0.001). Additionally, gaining of knowledge about PIH and GDM in the intervention group, compared to the control group, was significantly higher (p=0.001). According to the reported lab values, there were no differences in blood pressure and glucose levels between the two groups. Three cases of emergency childbirth due to PIH were reported in the control group versus one in the intervention group. Ninety three percent of the pregnant women who used CDA reported quicker identification of the alarming symptoms. Use of the CDA did not cause an increase in the anxiety level in the intervention group (before: 42.5±9.4 p=0.8, after: 42.2±9.2 p=0.9) compared to the control group (before: 42.1±11.3 p=0.8, after: 42.5±11.3 p=0.9).

Conclusion
CDAs could potentially improve self-efficacy and knowledge of pregnant women about their health status. As more ambulatory care settings migrate from hospitals to community care centers, informatics tools such as self-triage CDAs, can play an important role in such transitions.

References
Automating Identification of OEF/OIF Veterans Diagnosed with ALS

Jennifer M. Aucoin, MD, MPH1, Shon G. Michael, MD1, Amy C. Justice, MD, MSCE, PhD1,2, Perry L. Miller, MD, PhD1,2, Cynthia A. Brandt, MD, MPH1,2, Huned S. Patwa, MD1,3
1Connecticut VA Health Care System, West Haven, Connecticut, USA; 2Yale Center for Medical Informatics, New Haven, CT, USA; 3Yale University School of Medicine, New Haven, CT, USA

Introduction
Amyotrophic lateral sclerosis (ALS) is a progressive, fatal disease involving rapid degeneration and death of motor neurons. Family history, increasing age, and male gender are known risk factors for ALS, but numerous other factors have also been proposed1, including Persian Gulf military deployment2,3. This study endeavors to look at the feasibility of automating the identification of ALS patients in the Veterans Administration electronic record in order to allow for accurate measurement during an ALS-care quality improvement initiative.

Methods
The study sample consisted of a cohort of Operation Enduring Freedom (OEF) and Operation Iraqi Freedom (OIF) veterans in the Connecticut VA health system electronic medical record. Charts were queried for any existence of the ICD-9 code for ALS. The resulting charts were reviewed by a neurologist to confirm the diagnosis of ALS and to see whether the criteria established by the U.S. Agency for Toxic Substances and Disease Registry (ATSDR)4 were met.

Results
The results of chart review by a neurologist are shown in Table 1. Over 20 percent of charts containing the ICD-9 code were found to contain no apparent diagnosis of ALS. The majority appeared to be instances of miscoding, and only two of the patients met the ATSDR criteria.

Table 1. Classification of patients with at least one ICD-9 code of ALS (335.20).

<table>
<thead>
<tr>
<th>Classification</th>
<th>Percentage of patients</th>
<th>Number of patients</th>
<th>Patients meeting ATSDR criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>At least one ICD-9 code of 335.20</td>
<td>100%</td>
<td>149</td>
<td></td>
</tr>
<tr>
<td>No verifiable diagnosis of ALS</td>
<td>20.8%</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Apparent miscoding in ancillary clinic visit</td>
<td>11.4%</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Neurologic workup negative or other diagnosis</td>
<td>4.7%</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Patient report of outside diagnosis not confirmed</td>
<td>2.0%</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>No notes (blank chart)</td>
<td>2.0%</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Identified as duplicate</td>
<td>0.7%</td>
<td>1</td>
<td>NA</td>
</tr>
</tbody>
</table>

Conclusion
Previous attempts to identify veterans with ALS for inclusion in an ALS database have shown that ICD-9 code alone is insufficient criteria for confirming the diagnosis2. The criteria established by the ATSDR are reasonably accurate in identifying patients with a diagnosis of ALS. Patients who met criteria but were found not to have ALS were both prescribed Riluzole in response to previous prescription by outside providers.

References

1404
Combining indirect and direct methods to enhance research of online health communities

Uba Backonja, PhD RN¹, Albert Park, PhD¹, Andrea Hartzler, PhD²,
Megan N. Taylor, BS¹, Troy Griffith, BS¹, Wanda Pratt, PhD¹
¹University of Washington, Seattle, WA; ²Group Health Research Institute, Seattle, WA

Abstract

Although online health communities offer opportunities to characterize members’ communication and needs, researchers often use only indirect methods, limiting their inquiry. We discuss how our study combined indirect text processing and direct interviews with individuals who had cancer experience. Our approach allowed us to gain rich insights for the design of online communities that may not have been gained through indirect methods alone.

Introduction

When studying online health communities, researchers typically use indirect methods like online surveys¹ or large-scale text processing of existing posts², which have many benefits. Online surveys can be inexpensive and reach many people regardless of geography or time constraints. Analyzing existing data can be inexpensive and achieved in large volumes. Although these indirect methods unveil what individuals do in online health communities, they may not reveal a deeper understanding of how and why they engage in these communities. Also indirect methods may not reach non-users or members who do not contribute. Adding direct methods of inquiry, such as interviews, could overcome these limitations by providing nuanced information about members’ perceptions or behavioral motivations. This information can be invaluable when improving online health communities.

Utilizing both indirect and direct methods in online cancer community research

To understand how and why those with cancer experiences could use an online health community profile to select peer mentors, we conducted a study using online posts (indirect) and in-person interviews (direct). After receiving approval from our institutional review board, we automatically processed posts from an existing online cancer community to populate key elements of mentor profiles. We created paper mockups of profiles that described health interests and characteristics of potential mentors. Next we recruited patients and caregivers with cancer experience via fliers and cancer support groups. We conducted in-person interviews to gain an understanding of how and why participants preferred certain mentor profiles as they sorted and ranked profiles. Interviewers used open-ended questions to guide discussion and allow participants to converse freely and tell stories, providing contextual and emotional information to support ranking preferences. Participants thought out loud and took notes on paper mockups to elicit insights into their decision-making behind rankings. By combining direct and indirect methods in this online health community research, we had many opportunities to learn about individuals’ perceived challenges and benefits of joining online health communities and reasoning behind information they sought in profiles.

Conclusion

Using indirect and direct methods in online health community research provides nuanced understandings not typically obtained only via indirect methods. Interviews revealed informational and emotional decision-making of how and why participants chose mentors. Text processing allowed us to create profile mockups, providing realistic interview prompts. This information is important—it links to cancer patients’ online behavior³ but may be missed using indirect methods. Using indirect and direct methods provides invaluable insights about what, how, and why users engage with online health communities, which carry rich implications for online health community research.

References

2. Wen M, & Rose C. Understanding participant behavior trajectories in online health support groups using automatic extraction methods. GROUP ’12 2012;179-1883.
Learning Useful Abstractions from the Web

Abhishek Bafna(abafna@umich.edu)  Jenna Wiens(wiensj@umich.edu)

The successful application of machine learning to electronic medical records typically turns on the construction of an appropriate feature vector. Defining abstractions to create high-level, low-dimensional feature vectors can help in identifying clinically meaningful similarities among patients especially when the number of training examples is limited. In this work, we compare conventional unsupervised dimensionality reduction techniques (e.g., principal component analysis (PCA)) to new approaches that leverage a large corpus of freely available expert knowledge in unstructured form from the Web. As a case study, we consider the task of learning useful abstractions for patient medications.

We begin with a list of \( d \) medications and retrieve a corpus of text documents from Wikipedia, where each document corresponds to a medication. Next, we apply latent Dirichlet allocation (LDA) to this corpus [1]. This results in a representation where each medication (i.e., document) is associated with a distribution over \( k \) topics (\( k < d \)). One can think of the learned topics as classes of drugs, allowing the identification of similar drugs based on the topic distribution.

Given a cohort of patients and their medication orders, we propose a method for representing each patient in a \( k \) dimensional feature space and for computing a patient similarity/kernel matrix. We employ the earth mover’s distance (EMD) when building the matrix [2]. In contrast to the Euclidean distance, the EMD allows one to incorporate the underlying distances between topics.

Table 1: The area under the receiver operating characteristic curve (and the inter-quartile ranges) for the test set when 3% of the total training data (i.e., 500 samples) are used.

<table>
<thead>
<tr>
<th>Approach</th>
<th>30-day Mortality</th>
<th>Hosp. Readmission</th>
<th>ICU Readmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d )-Categories</td>
<td>0.61 (0.61-0.65)</td>
<td>0.67 (0.66-0.68)</td>
<td>0.57 (0.57-0.59)</td>
</tr>
<tr>
<td>PCA</td>
<td>0.63 (0.62-0.64)</td>
<td>0.60 (0.60-0.61)</td>
<td>0.59 (0.58-0.60)</td>
</tr>
<tr>
<td>LDA</td>
<td>0.64 (0.64-0.65)</td>
<td>0.61 (0.61-0.63)</td>
<td>0.61 (0.60-0.62)</td>
</tr>
<tr>
<td>LDA+EMD</td>
<td>0.65 (0.64-0.67)</td>
<td>0.63 (0.62-0.63)</td>
<td>0.62 (0.62-0.64)</td>
</tr>
</tbody>
</table>

Assuming patients with similar outcomes receive similar medications, an improvement in our ability to identify similar medications should translate to an improvement in our ability to identify similar patients. Thus, through a series of experiments, we measure the utility of the learned abstractions by considering three different clinical classification tasks: mortality within 30 days of discharge, readmission to the hospital and ICU after discharge and readmission to the ICU during the same hospital visit. We consider over 25,000 patient admissions from the MIMICII database [3]. After minimal preprocessing, we identified 2,285 “unique” medications. Applying the methods described above, we first mapped each medication to a vector of 60 topics. Next, we learned an EMD-kernel SVM classifier, for each outcome. In our experiments, we repeatedly divided the data by into training and test sets (70/30 split). We also varied the percentage of training data used from 1% to 100%. In addition to the proposed approach LDA+EMD, we considered three additional approaches for comparison:

- \( d \)-Categories - each patient is represented in the original \( d \)-dimensional feature space,
- PCA - uses principal component analysis and a kernel based on the Euclidean distance, and
- LDA - uses LDA analysis and a kernel based on the Euclidean distance.

Figure 1 shows the resulting test performance for the third task (ICU readmission). We note that when the number of training examples is limited (e.g., < 10%) the LDA+EMD model outperforms the three baselines. Table 1 displays similar trends for the other two prediction tasks. These results, support the hypothesis that one can learn useful abstractions efficiently by leveraging the large quantity of unstructured medical knowledge available on the Web.

Data Error Transparency in Health Information Exchange
Gina B. Baker, MSN1, Shan He, PhD1, Darren K. Mann1 Pallavi Ranade-Kharkar, MS1, 2, Sidney N. Thornton, PhD1, 2
1Intermountain Healthcare, Salt Lake City, UT; 2University of Utah, Salt Lake City, UT

Abstract
Health Information Exchange (HIE) systems complicate the current electronic medical record (EMR) structure and may perpetuate data errors as information is passed between healthcare providers. The longitudinal compilation of clinical data including the incorporation of inbound HIE data can form a baseline from which quality deviations can be assessed. Select types of data variability can be characterized and scored by logical processing. Alerting patients and providers to these possible errors in their EMR could potentially lead to decrease in adverse events and may improve data quality of transmitted clinical information.

Introduction
Data errors exist in the patient record even with EMR use. Several studies have documented the monetary and safety consequences of these errors; one study cites medication error rates as high as 37%1,2. Data errors of all types are likely to occur at transitions of care such as discharge to home or transfer to another hospital2. HIE systems are facilitating the information needs at transitions of care but may also inadvertently propagate the data errors. Some data errors are not detected until after they are transmitted to an external organization or a patient-facing interface. Weir and Nebeker suggest that these data errors lead to a lack of trust in the facility EMR3. Organizational risk and reputation may suffer when the data errors are detected outside the disclosing organization. Intermountain Healthcare developed screening rules for detecting data errors from externally acquired data sources including HIE data transmissions. Similar rules apply for outbound data to identify possible errors for the outside consumer.

Methods
Outbound data was evaluated for obvious and contextual errors specific to the patient or their encounter. Evaluation of the patient encounter was necessary in some areas to evaluate for patient-specific data such as body mass index. Obvious errors included limit checks for type mismatch, insufficient attribution for the data model or physiological incompatibility. A scoring system based on accuracy, completeness and clinical context was introduced to categorize the seriousness of the data error. Certain categories of data can be suppressed or filtered from the outgoing data package. For example, rules may be configured to filter aged or irrelevant data. Other data types may be flagged with cautionary guidance for future reconciliation workflows. Contextual or higher scoring data in question can be prioritized for review by the receiving provider or patient through their respective workflow tools.

Results and Discussion
Evaluation of the patient specific data from longitudinal patient record can help determine data that might be inconsistent or in conflict. Guidance and severity scoring can help alert the patient or provider receiving the data for the need to review and reconcile, thus improving the accuracy of the medical record. Feedback may be provided to local clinical and quality departments for systematic improvement.

Conclusions
Clinical data moving between organizations can be scrutinized and scored for integrity. Automated and select manual workflows can be enabled to control data quality. Metadata for data quality concerns can be shared among organizations and can improve the community view of a patient record. Transparency into data quality may translate into improved patient care with reduced treatment errors, and may increase patient engagement.

References
Patient-Centered Postoperative Wound Surveillance Using Smartphone Digital Photographs
Kate L. Baker, MA¹; Andrea N. Mahnke, MS¹; Sara Fernandes-Taylor, PhD²; Jason T. Wiseman, MD²; K. Craig Kent, MD²
¹Marshfield Clinic Research Foundation Biomedical Informatics Research Center, Marshfield, WI
²Department of Surgery, University of Wisconsin-Madison, Madison, WI

Background
Surgical site infection (SSI) is the most common nosocomial infection in surgical patients and accounts for 38% of post-operative complications¹. If diagnosed at an early stage, SSI can be treated in the outpatient setting with oral antibiotics and wound care, precluding the need for readmission, intravenous antibiotics, and reintervention. Patients rarely recognize early stage wound infections and often require intensive treatment and rehospitalization². It also produces increased health care costs and is the leading cause of unplanned, potentially preventable hospital readmissions for surgical patients³. Although interventions have been designed and employed to prevent SSI, the focus has been on the surgical procedure itself and the immediate postoperative inpatient stay¹. A large proportion of wound infections, however, develop following hospital discharge and prior to routine follow-up¹. The fact that SSI develops or progresses in the outpatient setting makes transitional care coordination an important focus in the effective management of SSI. Patient-centered technological interventions to improve transitional care for surgical patients are necessary to stem the burden of SSI and readmissions.

Objective
We designed and tested a patient-centered, outpatient wound surveillance program using smartphone digital photography to promote early recognition of SSI following discharge with the goals of (1) empowering patients to partner with their surgeons in monitoring their SSI symptoms and overall postoperative recovery, (2) diagnosing SSI at an early stage, enabling outpatient management, and (3) preventing hospital readmission and the serious morbidity and mortality associated with wound complications. We focused on vascular surgery because of the high risk of SSI and readmission among these patients.

Methods
We designed a patient-centered smartphone app that enables patients to transmit wound photos and symptom information to the vascular service. Images are uploaded to the PACScan image system, permitting side-by-side review of chronological images with linkage to the medical record. In collaboration with geriatricians and community/patient research advisors, we designed a patient-centered training process to teach patients or caregivers to take and transmit photos of their wound using smartphones. We conducted focus groups with community/patient research advisors to assess ease of use of the app and effectiveness of the training process. We utilized print and picture based teaching materials, “teach-back” by the patients, and pre-discharge reinforcement with a test photo transmission. We undertook formal usability testing of the app on an inpatient vascular surgery service to evaluate (1) barriers to participation, (2) patient experience, (3) picture and information quality, and (4) successful information transmission and assimilation into PACScan. Screen and audio recordings were captured during the inpatient usability testing for analysis and identification of usability issues. Improvements were recommended and implemented iteratively.

Results & Conclusions
Data collection and analysis pending.

Supported by the Agency for Healthcare Research and Quality R21 1R21HS023395 R21 HS023395

References
Integrating Conceptual Models to Inform the Design of a Family Health Information Management System for Hispanic Dementia Caregivers

Suzanne Bakken, PhD, RN,1,2 Maribel R. Granja MPH, 1 Niurka Suero-Tejeda, MS, MA, CHES,1 Samantha B. Stonbraker, MPH, RN,1 Robert J. Lucero, PhD, MPH, RN3,4
1School of Nursing and 2Dept of Biomedical Informatics, Columbia University, New York, NY 3College of Nursing, University of Florida, Gainesville, FL 4VA HSR&D Center of Innovation on Disability and Rehabilitation Research, Gainesville, FL

Introduction

Little is known about the information and communication needs of family caregivers for Hispanics with dementia or the manner in which online tools may meet those needs. As part of the New York City Hispanic Dementia Caregiver Research Program (NHiRP), we are developing a Family Health Information Management System (FHIMS) through participatory design methods with caregivers. The purpose of this poster is to describe the development of a coding framework for the initial participatory design sessions.

Methods

We conducted 21 participatory designs (10 English and 11 Spanish) with a total of 28 participants. Audiotapes of the sessions were transcribed and coded in the language of the session. The goal of the qualitative data analysis was to understand the dimensions of caregiving as well as the information and communication needs of those who care for persons with dementia as a foundation for designing the FHIMS. To create a coding framework for the analysis, we integrated concepts from Coiera’s1 model of the communication task-information task continuum with a clinical adaptation2 of the Krikelas model of supplemental information seeking behavior.

Results

The coding framework includes five top-level concepts: Caregiver, Patient, Information Needs/Tasks, Communication Needs/Tasks, and Online Tools. Key concepts from Coiera that informed the coding framework were information task, communication task, information agents, communication agents, and common ground. The concepts of patient information, domain information, and logistical information were used to further define information needs and tasks. In terms of common ground, communication needs/tasks were categorized by those in the conversation with an assumption of variation in amount of common ground, e.g., caregiver to caregiver vs. caregiver to healthcare professional.

Conclusions

The integration of concepts form several models resulted in a coding framework that is being applied to inform the design of the FHIMS.

Acknowledgment

Funded by New York City Hispanic Dementia Caregiver Research Program. R01NR014430. Ms. Stonbraker is supported by T32NR013454, funded by National Institute for Nursing Research.

References


1409
Analysis of Workflows: How Workflows Interact with CPOE to Contribute to Errors

Eleanor J. Barone, MD¹, Anne Miller, PhD²

Vanderbilt University Department of Biomedical Informatics, Nashville, Tennessee¹
Center for Research and Innovation in Systems Safety (CRISS), Vanderbilt University Medical Center, Nashville, Tennessee²

Introduction

The introduction of electronic health records (EHR) and computer provider order entry (CPOE) have decreased medication errors but at the cost of disrupting workflows.¹ This project explored the possible reasons for wrong order errors by investigating the interaction between workflows and CPOE at a major medical center, and how those interactions may contribute to or mitigate wrong order errors.

Method

Vanderbilt University Medical Center implemented its in-house CPOE in 1994. From 7/25/13 to 7/25/14, Quality Management identified 9 incidents involving wrong patient orders with CPOE as a potential contributing factor. A human factors investigation was requested and a three stage investigation implemented: Stage 1 - researchers interviewed quality analysts regarding original findings; Stage 2 - we observed 6 residents and 3 attending physicians in the Emergency Department (ED); 4 residents and 2 attending physicians in the Medical Intensive Care Unit (MICU), and 2 attending physicians on the Hospitalist Service. During the observations, the providers were asked to explain what they were doing as they did it. Observations were recorded on paper. Stage 3 – Data was analyzed using the Functional Resonance Analysis Method (FRAM)² to model and interpret the observations. The FRAM models work functions as defined by their inputs and outputs, and as mediated by time, regulatory process, and resources. FRAM supports explorations of ‘why things go right or wrong’ and is helpful for identifying adaptive redundancies as well as vulnerabilities in sociotechnical systems.

Results

Stage 1: Based upon analyst interviews, two of the wrong orders were same last name errors. CPOE was the only common factor for all of the errors. Eight errors were variations on a patient receiving the wrong test or medication. A radiology report read into the wrong patient chart was the most benign error.

Stage 2: User interface design issues were observed to contribute to error by making patient identification difficult and error prone. Order set popup screens, for example, obscured patient identifying information, inconsistent patient identifier placement reduced cross-checking, and the use of medical record numbers to select patients enhanced the risk of wrong patient selection. In all settings, ED, MICU, and Hospitalist Service, redundant communication and checks between staff on each patient order, plan, and result were critical to minimizing the risk of wrong patient orders.

Stage 3: FRAM modeling showed different communication and information flow pathways in each setting. In the MICU, bedside team rounding supported patient identification and enhanced cross checks and follow-ups. Hospitalists in small interdisciplinary teams functioned as sole entry points and so closed information and feedback loops. The ED, however, relied heavily on each provider’s individual sense of responsibility for the patient’s outcome with very limited interdisciplinary communication or cross checking. Additionally, ED ordering tended to be a ‘ballistic’ process; once sent, erroneous orders were difficult to identify and information flow loops could not always be closed.

Discussion

No one solution is evident. Wrong patient orders are particularly insidious as they involve two victims (the patient who did and the patient who did not receive a needed order). The prominent display of patient information, including patient photographs may better support order cross-checking, but it may not identify patients not receiving orders. Association between orders and problem/diagnosis lists may help providers in assessing order appropriateness, and support wrong order identification earlier. However, social processes that enhance team work, communication and cross-checking enhance resilience overall.

References

Visualization of Clinical Decision Support Failures

Mujeeb Basit, MD\textsuperscript{1,3}, Adam Wright, PhD\textsuperscript{2,3,4}
\textsuperscript{1}Beth Israel Deaconess Medical Center, Boston, \textsuperscript{2}Brigham and Women’s Hospital, Boston, MA
\textsuperscript{3}Harvard Medical School, Boston, MA, \textsuperscript{4}Partners HealthCare, Boston

**Introduction:** Evidence suggests that CDS can improve health care quality, safety, and effectiveness (1). Prior work has demonstrated multiple failures and potential harm from CDS improperly executing (2). Visualization of this complex data may help CDS implementers identify potential malfunctions of CDS rules. The visualization complements automatic methods for detecting CDS failures and identifying the key characteristics of failure. Visualizing temporal data can be complicated by natural variations in daily rates, including holidays, periodicity (weekends and seasons) and secular trends

**Methods:** We extracted CDS rule firing data from the EHR used at Brigham and Women’s Hospital, and visualized the data using the ggplot2 R package. Two rules “CAD and no ASA” and “Adult Seasonal Influenza” were selected for demonstration. Firing rates were adjusted for visit volume (using notes written as a proxy).

**Results:** Figure 1 shows the preliminary and final versions of the date visualization in a calendar format. The visualization clearly displays two known malfunctions in the CAD/ASA alert the first from 2009 being a system-wide spike caused by a system update and the second from 2012 caused by a corruption of the drug class manager resulting in excessive execution. The third graph shows the seasonal data and changes over the years. Each year the start and magnitude of the execution has changed. There were no anomalies noted in this data.

The enrichment of the circular graph focused on seasonal information.

**Discussion:** CDS implementation and monitoring is critical in maintaining patient safety and clinical trust. There exist multiple potential causes of failures of these systems. Clinicians have come to depend on these systems therefore identification of failures is critical. Visualization of this complex dataset is helpful to identify the pattern of failure and facilitate with correction. In future work, we plan to expand our system to optimize for variation in seasonal trends and provide linkage to a real-time dashboard for detecting anomalies before they lead to widespread patient harm.

**Figure 1:** Firing rate of “CAD and no ASA” alert at BWH over a five-year period demonstrated with various filters and the final version with annotation and enrichment. The third panel shows fire rate for “Adult Seasonal Influenza” alert demonstrating complexity of seasonal data.

**References**
A Faceted-Search Mobile App
for Matching Cancer Patients to Targeted Therapy Clinical Trials

Thomas P. Bechtold, M.B.A.1, Elly J. Cohen, Ph.D.2, Susan L. Colen2,
Laura J. Esserman, M.D., M.B.A.2
1QuantumLeap Healthcare Collaborative, San Francisco, CA; 2UCSF Department of Surgery,
San Francisco, CA

Introduction

Difficulties in patient recruitment often delay timely completion of cancer clinical trials. This can be even more challenging for targeted therapy trials where eligibility is restricted to patients who have tumors with a specific molecular signature. Since most patients who participate in trials do so at the suggestion of their providers, the availability of an easy-to-use mobile app that matches patients to trials by tumor stage and biomarker status could facilitate recruitment at the point-of-care. This is consistent with the growing use of healthcare apps by providers for decision support and patient communication during the clinical encounter1.

The Problem: Finding Clinical Trials at the Point-of-Care

Most clinical trials search tools are web-based and use either keyword searches or lengthy questionnaires. Keyword searches require the user to know a suitable keyword, and yield significant false-positive results. Questionnaire-based searches impose the up-front burden of entering a detailed health history. Furthermore, these resources do not perform accurate and up-to-date matching based on biomarker status.

Our Solution: Faceted Search with Biomarkers

We have developed a new service (CTMatch, ClinicalTrialsMatch.org) that finds cancer clinical trials with faceted search. Faceted search has been demonstrated to improve complex searching, especially with dynamic taxonomies. CTMatch prompts the user with facets specific to their indication. Given the proliferation of targeted therapy trials, it features accurate matching to biomarker test results, made possible by an innovative and scalable curation tool (Table 1, Figure 1). For example, a breast cancer patient seeking a trial will be prompted for breast-specific biomarkers ER and HER2/neu. The taxonomy - demographics, cancer stage and biomarker status – was chosen for its discriminative power.

Table 1. Biomarkers used in CTMatch faceted search

<table>
<thead>
<tr>
<th>ALK</th>
<th>KRAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR</td>
<td>B-RAF</td>
</tr>
<tr>
<td>Estrogen/Progesterone Receptors</td>
<td>BRCA1</td>
</tr>
<tr>
<td>HER2/neu Receptor</td>
<td>BRCA2</td>
</tr>
</tbody>
</table>

This matching service is distributed as Apple and Android mobile apps. It can also be inserted on a web page or accessed as a web service from clinical decision support applications.

Figure 1. Percentage of US cancer trials (n=3865) with eligibility criteria involving biomarker status

Evaluation

Usability testing will validate the efficacy of the app.

Conclusion

CTMatch is a new mobile app that uses a faceted search design, including stage and biomarker status, to find cancer clinical trials. In the future, the ability to capture and track critical care elements at the point of care will enable true automation of trial matching, making CTMatch even more convenient and integral to care.

References

Social Network Analysis: Data Collection Challenges and Solutions

Marge Benham-Hutchins, PhD, RN¹
Barbara Brewer, PhD, RN, FAAN²

¹University of Texas, Austin, TX ²University of Arizona, Tucson, AZ

Patient safety and quality initiatives have emphasized the importance of health information technology (HIT) design that supports communication and collaboration between providers across the continuum of care. Health organization decision makers, researchers, and designers of HIT have recognized compatibility issues between HIT and multi-professional communication processes resulting in problems integrating HIT into existing provider workflow. These problems have been associated with new types of medical error that are influenced or caused by HIT. Solutions depend on a closer examination of how providers access information and communicate with other providers in the provision of collaborative patient care. This has led to an interest in Social Network Analysis research methods.

Network analysis methods acknowledge that social networks are composed of individuals and the relationships between them. These relationships can be analyzed based on specific characteristics (e.g., frequency or quality of communication). In the complex health care environment relationships between providers may or may not be reciprocated, and individual providers can be connected to more than one other person or group. Social Network Analysis methods have been used successfully to explore the impact of information technology on provider communication ¹, to examine communication in emergency departments ², to identify information exchange networks in the ICU ³, and to describe medication advice-seeking in a renal unit ⁴.

 Nonetheless, designing and implementing a methodologically sound Social Network Analysis data collection process in the healthcare environment has proven to be challenging for both new and experienced researchers. This poster will provide preliminary results from a systematic review of the literature that examines common challenges and solutions associated with collecting network analysis data in the health care environment. Preliminary information on data collection methods, issues and solutions from the author’s ongoing longitudinal network analysis study will be included.

References
Using the Adverse Event Reporting System: Can Analysis be Streamlined by Text Processing?
Andrea Benin, MD1,2, Samah Jamal Fodeh, PhD1, Michele Koss, RN2, Kyle Lee2, Perry Miller, MD, PhD1, Cynthia Brandt, MD MPH1

(1) Yale University School of Medicine, Yale University, New Haven, CT, 06520
(2) Connecticut Children’s Medical Center, Hartford, CT, 06106

Background: Existing, widely used, software systems for reporting adverse events are limited in how they handle the unstructured text entered by frontline staff. This limitation means that use of this potentially rich information depends on extensive manual sorting and collation of electronic data, which is resource-intensive and prone to variable results.

Aim: The aim of this study was to develop an electronic approach to processing the text in medical event-reports that is reliable enough so that it can be used to improve patient-safety in a less manually resource-intensive fashion.

Methods: We categorized event-reports from February 2012-January 2014 at our free-standing children’s hospital using 3 different approaches: (1) manual review by experienced clinicians (“reference standard”), (2) rule-based query, and (3) semi-supervised machine learning: a hybrid method that combines F-score feature selection with Laplacian support vector machine. The 2 computerized approaches were compared to the “reference standard” and to the “usual approach” (categorization by the frontline personnel in the vendor-software tool). We focused on 5 use cases that were chosen specifically for their daily importance to patient-safety; 3 were known to be near impossible to identify using only the vendor-based query-tool. Being able to electronically identify the reports relevant for these chosen use cases is meaningful for improving patient-safety. All 9164 reports were annotated for the 3 use cases that were not readily amenable to direct query using the vendor-tool and “usual approach.” A subset (437) of the event-reports was also annotated for the 2 use cases that are part of standard taxonomies and known to be readily able to be queried using the vendor-tool and the “usual approach.”

Results: Both electronic approaches perform better than the “usual approach” and well enough to be used in practical applications to improve patient-safety.

<table>
<thead>
<tr>
<th>Use Cases: Category of Event (number of event reports in annotated set applied for each category)</th>
<th>Manual Annotator [“reference standard”], n</th>
<th>Rule-Based Query, n (% agreement ,both positive and negative, Kappa)</th>
<th>Machine Learning, compared to set of positive annotations, n (% agreement)</th>
<th>Frontline staff entry [“usual approach”], n (% agreement, both positive and negative, Kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight-Based Error (9164)</td>
<td>102</td>
<td>203 (99%, 0.62)</td>
<td>79/85 (93%)</td>
<td>Not querable in standard taxonomy</td>
</tr>
<tr>
<td>Error in Identification of Patient (9164)</td>
<td>1517</td>
<td>3141 (81%, 0.52)</td>
<td>1013/1102 (92%)</td>
<td>350 (87%, 0.33)</td>
</tr>
<tr>
<td>Activation of Medical Emergency Team (9164)</td>
<td>74</td>
<td>85 (99.8%, 0.89)</td>
<td>Not attempted</td>
<td>Not querable in standard taxonomy</td>
</tr>
<tr>
<td>Infiltrate of an Intravenous Catheter (437)</td>
<td>26</td>
<td>26 (99.5%, 0.96)</td>
<td>Not attempted</td>
<td>10 (96%, 0.54)</td>
</tr>
<tr>
<td>Fall or Drop (437)</td>
<td>23</td>
<td>17 (99%, 0.84)</td>
<td>Not attempted</td>
<td>13 (97%, 0.65)</td>
</tr>
</tbody>
</table>

Conclusions: Electronic approaches to streamlining the use of free-text entered into an adverse event report system are feasible and can be extremely valuable in categorizing and sorting this important data for use in improving patient-safety. In addition, the common taxonomy used by this widespread software vendor does not readily support accessing all aspects of the information that is needed to improve clinical care; alternate, automatable query-methods offer important flexibility to the users of this data hoping to improve patient-safety.
Demographic Factors Associated with Differences between New York Inpatient Medicare Charges and Payments for (DRG 065) Stroke

Barbara A. Berkovich, MA, E. Shannon Stearman, BA, Craig W. Johnson, PhD
The University of Texas Health Science Center at Houston (UTHealth)
School of Biomedical Informatics, Houston, Texas

Introduction
Stroke results in morbidity and mortality and is costly to treat. Hospital costs have exceeded Medicare reimbursements for ischemic stroke. In 2007, additional levels for complication and comorbidity were added to the Diagnosis Related Groups (DRGs) of the Centers for Medicare Services (CMS) to establish diagnosis-based payments. Research to establish predictable cost drivers for hospitalizations continues. Using IRS New York State 2012 tax returns, we determined whether differences between Medicare charges and payments for DRG 065 (intracranial hemorrhage or cerebral infarction with a major complication or comorbidity) were associated with rural versus urban locations, total income, and unemployed households, as determined by zip code.

Methodology
Dataset 1 included all New York State records for hospital-specific charges per discharge using the Medicare Severity Diagnosis Related Group (MS-DRG) for Fiscal Years 2011 and 2012. Dataset 2 is from the IRS website Statistics of Income Tax Stats –Historic Table 2, which is a record of tax returns (Form 1040) from the IRS Individual Master File system from January 1, 2013 to December 31, 2013. We analyzed data from 91 urban and 25 rural providers in 103 unique zip codes. For each zip code, we integrated charges and payments from Dataset 1 with Dataset 2’s amount of income tax, number receiving unemployment, and whether farm income was reported. The latter three variables operationalized total income, unemployed households, and rural versus urban locations, respectively. We eliminated four New York State zip codes (10802, 11794, 13503, 14642) due to missing data. Due to unequal sample sizes and a nonnormal distribution, parametric and nonparametric (using Puri-Sen L statistic) univariate analyses of covariance employed SPSS 22 UNIANOVA to compare Medicare charge versus payment differences for rural and urban locations. Total income and unemployed households served as covariates.

Results and Conclusion
Rural versus urban locations (L = 10.02, p = .005; F = 8.50, p = .004) and total amount of tax (L = 12.85, p = .001; F = 11.82, p = .004) were significantly related to the Medicare charge versus payment difference. Urban location covariate-adjusted Medicare charge versus payment difference was $8,411.37 higher than rural, primarily attributable to $10,745.37 higher average charges in urban locations (Figure 1). Unemployed households were not statistically significant. Across all zip codes, and independent of urban versus rural location, the charge versus payment difference increased as a 1% (p = .001) multiple of the amount of tax. In other words, the wealthier the zip code the greater the charge versus payment difference.

References
GeneLab: NASA’s Open Access, Collaborative Platform for Systems Biology and Space Medicine

Daniel C. Berrios, MD MPH PhD1,2, Terri G. Thompson, PhD2, Homer W. Fogle, MS2, Jon C. Rask, MS2, Joseph C. Coughlan, PhD2
1University of California, Santa Cruz, CA; 2NASA Ames Research Center, Moffett Field, CA

Purpose

NASA’s mission includes furthering our understanding of biological systems through space-based research in order to improve life on earth and to enable the human exploration of space. To achieve these goals, NASA is investing in GeneLab1 (http://genelab.nasa.gov/), a multi-year effort to maximize utilization of the limited resources to conduct biological and medical research in space, principally aboard the International Space Station (ISS). Through the GeneLab project, researchers will include ground-based high-throughput genomic, transcriptomic, proteomic or other “omics” analyses as part of their experiments conducted on the ISS. The data from these analyses will be stored in the GeneLab Data Systems (GLDS), an information system currently being developed for the project. GeneLab intends to support “open science” research on the housed data sets, creating a multiplier effect on the science return from these experiments; the GLDS will serve all omics data without restrictions to the public. The system will ultimately include a biocomputation platform with collaborative science capabilities, to enable the discovery and validation of molecular networks that are influenced by space conditions.

Design

NASA has chosen a phased capability implementation for the GLDS. The initial phase of the GLDS development effort emphasizes capabilities for the submission, curation, search, and retrieval of omics data sets. Important design considerations have included 1) leveraging existing systems and systems components to deploy Phase 1 capabilities expediently; 2) determination of optimal data set curation procedures including metadata representation and generation and quality control procedures; and 3) balancing GLDS accessibility with user engagement and usage tracking capabilities. Phase 2 will focus on interoperability and supporting federation of GLDS-housed data sets with externally-curated data in data set search, retrieval, and annotation functions. In Phase 3 we will develop a platform for computational biologists to execute and collaboratively develop analysis tools, integrate results of analyses with those of collaborating researchers, annotate data sets with interpretations, and share insights and hypotheses, building a knowledge base relevant to space biology and medicine.

Challenges

The GeneLab project and the GLDS face several challenges, including:

- Numerous and rapidly evolving standards for omics data, metadata, protocols and analyses
- Representing the unique aspects of space biology experiments as metadata and related data
- Coalescing access to a wide range of biocomputation tools, overcoming lack of interoperability
- Incentivizing engagement by the space biology/medicine research community

Conclusion

High-throughput systems for analyzing genomes and transcriptomes are generating vast amounts of biological data. Integrated analyses of these data with proteomic, metabolomic, physiologic, and phenotypic information holds the promise of rapid elucidation of complex molecular pathways, and of huge advancements in biological understanding and space medicine. Data Systems like the nascent GLDS are challenged with the representation, organization and integration of these highly complex data sets, and with providing researchers with the tools and environments they require for maximum utilization of the data in their analyses.

References

Introduction

Electronic health records (EHRs) have emerged as a promising tool to improve health care quality. Significant federal investments are being made to encourage EHR adoption through incentive programs, however nursing homes (NHs) are excluded from these initiatives. Facing substantial barriers to EHR adoption, NHs appear to be lagging behind in EHR adoption rates, compared to other health care settings. However, research systematically assessing rates of EHR adoption in NHs on a national level is lacking. To that end, this study aimed to assess EHR adoption in a national sample of NHs.

Methods

As part of a larger mixed methods study (R01NR013687) to explore infection control and prevention in NHs, cross-sectional 34-item surveys were conducted in a randomized national sample of NHs. Questions were included in the survey to assess whether EHRs had been implemented in the NHs and what year. Additionally, one question asked about tools and resources used in the facility to help make clinical decisions, including clinical decision support systems (CDSS). Descriptive analyses were performed using the SAS 9.3 statistical software.

Results

Completed surveys were received from 990 NHs. Half of participating NHs (51%) reported having implemented EHRs. Of those, 76% had implemented EHRs in 2010 or later, and 35% in 2013 or later. Only 11% of the NHs where EHRs had been implemented reported that CDSS was used in the facility to help make clinical decisions.

Conclusion

The findings of this study indicate that EHRs are still a relatively new phenomenon in NHs and adoption still lags behind acute and primary care settings. More work is needed to increase adoption. Furthermore, to achieve meaningful use, health care settings are required to implement CDSS but in this study only 11% of NHs using EHRs reported using such systems. Future research should go beyond examining adoption rates to assessing the level of implementation and use of EHRs in NHs.

References

An Ontology-Driven Patient History Questionnaire System

Jonathan P. Bona, Ph.D., Gunther Kohn, B.S., Alan Ruttenberg, M.S.
University at Buffalo School of Dental Medicine, Buffalo, NY,

Our group at the University at Buffalo School of Dental Medicine is developing an ontology-driven system for collecting, recording, and managing patient history data. It consists of an HTML5-based web application backed by an RDF triple store database. The triple store is populated with explicit representations, in the Web Ontology Language (OWL), of patients and the entities - disorders, diagnoses, activities - that medical histories are about. The underlying ontology also directly represents the processes involved in collecting and recording this data, and the questions, answers, and other information artifacts involved. It links those artifacts to the entities in the world that they are about.

This standalone system interoperates with our clinic information management system, Picasso, as part of a pilot project to test the usefulness and effectiveness of the ontology-based approach. It replaces a series of paper questionnaires that are currently used to perform initial assessment of patients, and to gather both general history as well as oral and dental health history. Though it will be used initially in our dental clinics and include oral health questions along with general health history, the design is domain independent and driven by flexible ontology-based models. The underlying ontology will be publicly available for others to use and extend in similar projects.

As a dental school, our mission includes not just patient care but also education and research. This system is intended to support all parts of that mission. It is designed to collect and store patient history data in a way that preserves its meaning independent of any one use. By using explicit ontology-structured representations, we represent information about the patient's history independently of how it was produced (i.e., the particular user interface and business logic that currently make up different levels of the software). By representing that data natively in OWL and storing it in a triple store that supports reasoning, we make that data immediately available to logical inference and for authorized users and connected information systems to query using SPARQL.

We make extensive reuse of existing science-based ontologies associated with the OBO Foundry. These include the Ontology of Biomedical Investigations (OBI), the Information Artifact Ontology (IAO), the Ontology for General Medical Science (OGMS), and the Oral Health and Disease Ontology (OHD). Each of these includes terms relevant to our task that have clear English definitions, computable logical definitions, and links to a shared upper-level ontology, Basic Formal Ontology (BFO).

The system is designed for ease of use. Rather than just recreating a paper questionnaire in software as a list of questions, we have included the ability to dynamically adjust which questions are shown based on answers to previous questions. The use of explicit representations of information artifacts makes it straightforward to model dependencies and other relations between questions and answers, and to use those to dynamically adjust the display.

Another of our goals is to completely capture provenance of any data that the system records. We explicitly represent the question-taking process, its participants (patients, clinicians), its sub-processes (asking questions), and its outputs (information artifacts about the patient’s health history). Every answer is linked to terms that identify the individual that recorded the answer, the exact time at which it was recorded, the patient who has this answer as part of their history, and so on. In our clinics, the person taking and recording a patient history is often a student working under the supervision of a faculty member who must formally approve the student's work at the end. This information - approval status of a history-taking session, identify of the approver, etc - is also captured.

A patient has one history that is made up of many statements about the patient's health, which are often recorded at different times, possibly in many different sessions of history-taking. Our system preserves the complete patient history over time. By default, a provider viewing or editing a patient's history sees only the latest answer to each question, but a provider in an authorized role may view the complete history for a particular question or group of questions, or even a snapshot of the complete history as it was at a particular time.

In conclusion: we are developing a medical history application based on a carefully constructed ontological model that represents not only the things that a patient history is about, but also elements of the history-taking process, including the questionnaire and its contents. The result is a flexible, easily queried knowledge base of patient histories with semantic representations that facilitate its use for research and in conjunction with other information systems.
AMIA 2015 POSTER SUBMISSION

TITLE: Development and implementation of a Floor Admit Reevaluation Alert (FARA) in a large academic emergency department.

AUTHORS:
- James Booth, MD, MSHI, University of Alabama at Birmingham
- Eta S. Berner, Ed.D, FACMI, University of Alabama at Birmingham
- Jorge Alsip, MD, University of Alabama at Birmingham

DESCRIPTION OF THE PROBLEM:
The severity of a patient’s illness determines whether a patient is admitted to a normal hospital floor bed or a higher acuity ICU bed. Emergency physicians are often the providers responsible for determining the severity of a patient’s illness and the patient’s admission order. The emergency department (ED) is a fast-paced error-prone environment and the severity of a patient’s illness may go unrecognized, resulting in a severely ill patient being admitted to a normal floor bed. This type of error is associated with increased mortality and increased rate of ICU transfers from the floor.

We sought to develop a clinical decision support (CDS) pop-up alert that informs emergency physicians, at the time of the admission order, of any significant vital sign abnormalities or lab abnormalities that are associated with increased severity of illness. By alerting the physician to dangerous vital sign and lab abnormalities, we hope to decrease the rate of severely ill patients being inappropriately admitted to a floor bed instead of an ICU bed. After a thorough literature review, it appears that this type of CDS alert is a new concept and we found little evidence that this type of CDS alert has been formally researched.

METHODS:
Using a combination of literature review and internal hospital data, we developed a list of 13 vital sign and lab abnormalities associated with increased patient mortality and ICU transfers from the floor. Using these 13 criteria, we developed a CDS pop-up alert that alerts the emergency physician if the physician attempts to admit a patient to a floor bed with one of the 13 criteria present. The alert was implemented at a single large academic medical center and 6 months of pre-implementation and 6 months of post-implementation patient data were collected and analyzed.

RESULTS:
From August 15, 2014 to February 15, 2015, there were 1633 alerts. 155 (9.5%) of the alerts resulted in a change from floor admission to ICU. 87 (5%) of the alerts resulted in a change from floor admission to a step-down unit (a level of acuity between a floor and ICU bed). The total number of ED admissions was 6.7% higher in the 6-month post-implementation period while the total number of ICU transfers from the floor, within the first 12-hours of floor admission, decreased by 4.7%. 16 patients required transfer from the floor to the ICU within 12-hours of their alerts being over-ruled/ignored in the ED.

CONCLUSIONS:
ED physicians use clinical judgment when determining the need for a high-acuity ICU rather than a lower-acuity floor bed when admitting patients to the hospital. Occasionally, ED physicians may not recognize the severity of a patient’s illness and admit a patient to a lower-acuity floor bed where the patient’s status may continue to deteriorate. Design and implementation of a CDS pop-up alert may decrease the rate of severely ill patients admitted to a floor bed instead of an ICU and consequently decrease the rate of ICU transfers from the floor. Future analysis of the 13 alert criteria will be used to further refine the alert to improve the sensitivity or specificity of the alert. Additionally, analysis of ICU transfers post-implementation of the alert will help further refine the criteria.
Integrating Electronic Health Record Competencies into Undergraduate Health Informatics Curricula: A Preliminary Qualitative Study

Elizabeth M. Borycki, RN, PhD¹, Janessa Griffith, MSc¹, Andre W. Kushniruk, MSc, PhD¹

¹University of Victoria, Victoria, British Columbia, Canada

Abstract

Integration of electronic health record (EHR) competency development throughout an undergraduate health informatics program is essential for students to understand how: (a) the EHR can be used by clinicians, (b) to procure an EHR for specific user interface requirements and workflows, (c) to customize a system for local (i.e. health care system and clinical/organizational practice) setting, and (d) to design, develop and build an EHR.

Introduction

Integrating electronic health record (EHR) competencies into a four year undergraduate health informatics (HI) program can be difficult. Understanding what competencies can be integrated at what point in time is critical to students being able to understand the theory used in the design of the technology and how it is used, procured, customized to local settings and designed and developed for use². The development of EHR competencies for those studying HI needs to be considerate of how the technology is created, and how it fits with a local setting. Some of these EHR competencies are specific to developer (vendor) needs (e.g. those surrounding design). Others are specific to large health care organizations, clinics, and physician offices, where the emphasis of knowledge is placed on clinician use of the technology, procurement to fit the local healthcare setting and customization to improve user interface and workflow fit. Educational research in the area of HI has focused on how to technically provide students with access to EHRs in the classroom setting, the barriers to providing access to EHRs, and the impact of introducing the EHR into the classroom upon competency development. In this poster we present on the preliminary findings from a qualitative study investigating how EHRs could be integrated into a four year undergraduate HI program to facilitate student competency development.

Learning Objectives

Learners will be able to formulate a plan to integrate EHR competencies into HI programs and will be able to identify and develop a strategy for integrating vendor and health care organizational EHR competencies.

Methods

We conducted a qualitative study to investigate students’ perceptions about how EHR competencies could be integrated into an undergraduate program. Ten, third and fourth year students, were recruited to participate following completion of a course where they were introduced to EHRs for the purpose of developing competencies. At this point in the HI program third and fourth year students had completed two cooperative education experiences where each student has worked for two to three, four month blocks in an industry setting. Students were interviewed by telephone at a time and place that was convenient for them.

Results and Conclusions

Several key themes emerged. Student competencies could be subdivided into those that are required by vendors and those by health care organizations. Vendor EHR competencies involved the design and development of new EHR features and functions and interfacing technologies (e.g. mobile devices). Health care organizational competencies involved procurement, customization (user interface design, workflow) and implementation competencies. There is a need to integrate EHR competencies into HI programs in a manner that allows for competency development in the first year of study, followed by the development of increasingly more complex competencies such as those related to procurement, customization and implementation of the technology followed by design, development and building of new EHR features and functions and associated technologies.

References

Home-care Scheduling, Supervision and Security (HC-SSS): A status report

Maria Botsivaly, MSc, Pierros Vassileios, MSc, Marinis Michael, BSc,
Aris Tzavaras, PhD and Basile Spyropoulos, PhD
Technological Education Institute (TEI) of Athens, Athens, Greece

Abstract

It is the aim of this project to describe in brief the work in progress of a Home-care Scheduling, Supervision and Security (HC-SSS) approach, designed to support the continuity of health-care and the secure information management, for patients discharged from Hospital to Home-care.

Introduction

Health-care provision is moving towards home-care that is expected to provide for the upcoming demand, for patient-centered care, reducing at the same time the level of reimbursement for health-care providers.

Methods and Resources

The Scheduling-module allows for any discharging physician to assign, an appropriate set of home-care patients’ activities, provided by attendants, nurses etc. to specific Diagnoses. These activities’ profiles are adaptable and every physician is allowed to set up his own profile, create a new one or modify an existing one, in order to adapt it to each patient’s specific characteristics, needs and demands, on a completely web-enabled modular platform.

The Supervision-module aims to generate personal health-records from home, allowing hospitals, health-care providers and families to track and respond to critical behavioral and clinical patient data. The assigned homecare activities are displayed in the Supervision application in the form of instructions and reminders. Furthermore, the system allows for automatic and manual entry of patients’ parameters that should be monitored, according to homecare plan and it allows for real-time audio-visual communication and/or uploading of any sort of digital data.

The Security-module: A security-web manages all requests that need to be secured, to and from the recipient. The server is implemented using Microsoft ASP.NET Web API 2. The exposed API requires the use of https/SSL in order to allow for the execution of any action. We have created a digital certificate on our IIS Server that the user has to accept and after the acceptance, all calls that are being made are considered secured at the transmission layer. We also use anti-forgery techniques to prevent cross site request forgery (CSRF), like altering the query string on the browser. All sensitive requests contain a hash-value that has to be validated on the post. As an extra level of security, we have enriched our tokens with different "salt"-values on each "logical domain", preventing, thus, an attacker that manages to get hold of a valid token, to use it in another domain of the application. After the security server has successfully received a message, it strips the valid information and gives the control to an application server, which knows how to process the requests. Both servers can be hosted under the same IIS or even be incorporated in a single application.

Results

The system comprises of a home-care scheduling and supervision module, complying with the ASTM E2369 (CCR) and ISO 13606-1:2008 Standards and comprising adequate cryptographic algorithms’ data-protection.

Conclusion

Web-based techniques are ideal for home-care applications, since they allow for numerous distributed users, while reducing the maintenance task to a group of web-servers at the same time.
Reducing Healthcare Costs through Medical Recommendations

Lina Bouayad, PhD1,2, Balaji Padmanabhan, PhD1, Kaushal Chari, PhD1
1University of South Florida, Tampa, FL, 2Center of Innovation on Disability & Rehabilitation Research (CINDRR) James A. Haley Veterans Hospital, Tampa, FL

Abstract
In the US procedure and drug costs are often not visible by physicians, leading to high costs of care. Results of an experiment completed by physicians indicate that presenting alternative procedure recommendations with similar outcomes and lower costs may potentially help reduce overall healthcare costs. Interestingly, time pressure greatly affects the physicians’ use of such recommender systems.

Introduction
Healthcare costs have reached skyrocketing numbers in the US, which is now in the leading place when it comes to healthcare spending. Medical fraud clearly contributes to rising costs. In addition, misuse and underuse of medical procedures also contributes to elevated healthcare costs.

Particular to healthcare prices for the same procedures vary significantly depending on the paying party. In an example reported by the Wall Street Journal, the average charge for a joint-replacement surgery ranged from $5,300 in Ada, Okla., to $223,000 in Monterey Park, California1. It is important to note that providers are typically unaware of healthcare costs. Physicians are able to differentiate identify generic drugs within each drug type. However, they remain unacquainted with exact tests and drug costs1.

What if cost information can be presented to medical providers at the right time? In this work we design a medical recommender system that presents alternative prescription options that are 1) appropriate for the patient being consulted, and 2) are lower in cost compared with the physician’s initial selection. Because of the time involved in evaluating different recommendations (especially in emergency rooms), recommender systems have been either selectively used or completely removed2. Taking time pressure into account may therefore be important if it can increase acceptance of such systems.

Methods
We conducted an experiment involving medical providers in Florida. Any clinician with prescription privileges in Florida was able to participate in the study. All participants were presented different patient cases and asked to select the most appropriate prescription regimen. A between-subjects design was used to test the effect of cost on the use of the recommendations. Participants were placed into two different groups presented with varying recommendation costs. One group of participants was presented a list of all low cost recommendations, while the other group received recommendations of mixed costs. On the other hand, a within-subjects design was used to evaluate the impact of time pressure on the physicians’ influence by time pressure. In half of the cases, the participants were given a time constraint indicating high time pressure. Counterbalancing was used to control for any learning effect that might have occurred. Therefore, high-time pressure scenarios were either presented first or last depending on the group of the participant. Participants were randomly placed into four different groups.

Results
We were to recruit 40 physicians for the experiment. Mixed logistic regression, along with decision trees were used to test the significance of the hypotheses. Analysis results will be presented at the conference.

Conclusion
Recommender systems could be used in the medical domain to present cost information to clinicians, and thereby reduce overall healthcare costs. While designing such systems, our results show that it is also important to take into account the physician’s amount of time pressure experienced.

References
Variable Importance in Recursive Feature Selection with Random Forests for Mortality Prediction in ICU

Sabri Boughorbel, Rashid Al-Ali

Sidra Medical and Research Center, P.O. Box 26999, Doha, Qatar

Method: Feature selection is a research field in machine learning that investigates methods and algorithms for the selection of the most relevant variables [1]. We propose to combine model-dependent and model-independent (feature-based) variable importances in the feature ranking step. This approach introduces robustness to the feature selection procedure. The algorithm [2] describes the recursive feature selection steps [2]. The feature-based variable importance is obtained by computing the ROC AUC using each variable separately. Hence we can evaluate the importance of each feature for the prediction of ICU mortality. The model-based variable importance is computed using the fitted Random Forests model [3].

Algorithm 1: Feature selection with the modified variable importance.

1. Fit the Random Forests model with all features p
2. Estimate variable importance using proposed variable importance

for Each feature subset $F_i$
do
4. Fit Random Forests with $F_i$ features
5. Estimate model performance
6. Re-rank the features using proposed variable importance
7. Compute the performance profile for all sets $F_i$
8. Select the best variable subset $F_{opt}$

Results: For the validation of the algorithm, we used CinC’12 ICU data described in [4]. For each variable, first-order summary statistics are computed from the time-serie data namely mean, max, median, first and last values. In total 124 features are extracted. Table 1 shows the top selected features for the usual and proposed variable importance. Figure 1 shows performance changes with respect to the number of features.

<table>
<thead>
<tr>
<th>Variables (RF)</th>
<th>V.I.</th>
<th>Variables</th>
<th>V.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCS_max</td>
<td>11.24</td>
<td>GCS_max</td>
<td>92.90</td>
</tr>
<tr>
<td>GCS_mean</td>
<td>9.09</td>
<td>BUN_mean</td>
<td>82.15</td>
</tr>
<tr>
<td>Urine_median</td>
<td>8.51</td>
<td>BUN_median</td>
<td>79.93</td>
</tr>
<tr>
<td>GCS_median</td>
<td>8.20</td>
<td>BUN_min</td>
<td>78.61</td>
</tr>
<tr>
<td>Urine_mean</td>
<td>7.48</td>
<td>GCS_mean</td>
<td>76.40</td>
</tr>
</tbody>
</table>

Table 1: Top 5 selected features and corresponding variable importances (V.I.) based on Random Forests model (left two columns) and the proposed variable importance (right two columns). GSC is Glasgow Coma Score (3-15), Urine is the urine output (mL), BUN is blood urea nitrogen (mg/dL).

Discussion: The proposed modification of the feature selection procedure introduces an external factor to the variable importance computation. This factor has the advantage of stabilizing the feature ranking stage since it becomes less prune to model error and sensitivity to correlated features. The analysis of the theoretical effect of using model dependent and independent variable importance is an interesting research topic to further confirm and generalize the proposed approach.

References


Developing Intermediary Medication Phenotypes via Metabolomics Biotechnology Using an Electronic Health Record-Linked Biobank

Matthew K. Breitenstein, PhD\textsuperscript{1}, Richard M. Weinshilboum, MD\textsuperscript{1}, Jyotishman Pathak, PhD\textsuperscript{1}, Liewei Wang, MD, PhD\textsuperscript{1}, \textsuperscript{1}Mayo Clinic, Rochester, MN

Introduction and Background

Metformin is the first-line therapy for type 2 diabetes mellitus (T2DM). Emerging evidence implicates the potential repurposing of metformin for cancer prevention and treatment (PLOS One 2013;8(8):e71583). However, the details underlying the molecular mechanism of action for metformin are not fully understood. Specifically, the pharmacodynamics (PD) of metformin, the physiological and biochemical impact of metformin in the body, are not clearly understood (Pharmacogenomics 2014;15(4):529-539), which limit availability of intermediary phenotypes for metformin impact in cancer. We seek to utilize combined electronic health record (EHR) data with a linked biobank and metabolomics biotechnology to elucidate metformin PD and develop intermediary medication expression phenotypes that might enable further research regarding the potential repurposing of metformin for cancer treatment.

Materials and Methods

A cohort of T2DM Mayo Clinic Biobank patients with platelet poor plasma samples was identified using a validated phenotyping algorithm (JAMIA 2012;19(2):212-18). A total of 42 pre-defined metabolites representing the function of key biological processes were selected for metabolomic assays. The final cohort (n=600) submitted for metabolomics analysis had cases (metformin monotherapy exposure on sample date, T2DM) and controls (no metformin exposure on sample date, no T2DM) matched by gender and age. Our preliminary cohort (n=207) contained randomly selected samples that were processed and analyzed at the time of this abstract.

Results

In our preliminary analysis cohort 8 metabolites passed Bonferroni correction in the filter step and were selected for further analysis by a cumulative logit model (Table). These associations represent a preliminary metformin metabolomics expression profile and are the subject of active functional and systems pharmacology research.

<table>
<thead>
<tr>
<th>Metabolite name</th>
<th>1: Filter Step</th>
<th>2: Analysis Step</th>
<th>OR</th>
<th>p-val</th>
<th>LCL</th>
<th>HCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha-Aminoadipic-acid</td>
<td>0.259</td>
<td>0.0002</td>
<td>4.142</td>
<td>&lt;0.0001</td>
<td>2.210</td>
<td>7.763</td>
</tr>
<tr>
<td>Citrulline</td>
<td>-0.314</td>
<td>&lt;0.0001</td>
<td>0.199</td>
<td>&lt;0.0001</td>
<td>0.107</td>
<td>0.373</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>0.335</td>
<td>&lt;0.0001</td>
<td>8.598</td>
<td>&lt;0.0001</td>
<td>4.408</td>
<td>16.770</td>
</tr>
<tr>
<td>Alanine</td>
<td>0.421</td>
<td>&lt;0.0001</td>
<td>8.351</td>
<td>&lt;0.0001</td>
<td>4.319</td>
<td>16.145</td>
</tr>
<tr>
<td>Proline</td>
<td>0.306</td>
<td>&lt;0.0001</td>
<td>4.691</td>
<td>&lt;0.0001</td>
<td>2.518</td>
<td>8.737</td>
</tr>
<tr>
<td>Valine</td>
<td>0.287</td>
<td>&lt;0.0001</td>
<td>4.342</td>
<td>&lt;0.0001</td>
<td>2.329</td>
<td>8.095</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>0.297</td>
<td>&lt;0.0001</td>
<td>4.821</td>
<td>&lt;0.0001</td>
<td>2.575</td>
<td>9.028</td>
</tr>
<tr>
<td>Leucine</td>
<td>0.284</td>
<td>&lt;0.0001</td>
<td>4.533</td>
<td>&lt;0.0001</td>
<td>2.423</td>
<td>8.482</td>
</tr>
</tbody>
</table>

Bonferroni threshold p=.0012 *age, gender, batch adjusted

Conclusion

Our EHR-linked biobank metabolomics study identified a novel preliminary metabolomic profile of metformin exposure in a T2DM population. The metabolites in this profile have potential to serve as intermediary phenotypes in future research endeavors regarding metformin mechanism and/or repurposing for cancer treatment or prevention and will be further refined in ongoing functional validation. Further, this study represents a novel application of informatics phenotyping in an EHR-linked biobank that retrospectively identified a case-control cohort with utility for drug signature development. Finally, this study highlights the translational power of using EHR-based phenotypes to leverage a biobank to drive clinical and pharmacology discovery in the era of precision medicine.
The Internet: A Source of Near Real Time Infectious Disease Information

Juliana J. Brixey, PhD, MPH, RN¹, Catherine Pepper, MLIS, MPH¹, Jacqueline E. Brixey, MA²

¹University of Texas Health Science Center at Houston, Houston, TX; ²University of Texas at El Paso, El Paso, TX

Abstract
Event-based reporting of infectious diseases is an alternative to indicator-based reporting to identify potential cases and outbreaks. Social media have contributed epidemic intelligence to population health surveillance. A search of the peer-reviewed journal literature and of the Internet yielded a set of social media and web-based tools that have been used in event-based disease surveillance reporting.

Introduction
Primary care and emergency department (ED) clinicians can no longer wait days or weeks for infectious disease surveillance information. Global travel by air, sea, or high-speed rail has made the worldwide dispersal of an infectious disease within a few hours a reality. Public health agencies use both indicator-based and event-based surveillance reporting as accepted methods to establish epidemic intelligence¹. Indicator-based reporting entails the systematic routine collection of structured data from health care organizations, diagnostic laboratories, and health care providers to identify known cases or clusters of disease.¹ The lag time from reporting to data analysis and dissemination of findings has the potential to encumber the promptness in identifying an infectious disease outbreak. In contrast, event-based reporting does not rely exclusively on official sources of information, but is grounded in the rapid capture of information about events and rumors of illness or deaths from unstructured data sources such as the news media, news aggregators, Twitter, Facebook, blogs, online discussions, and Really Simple Syndication (RSS) to detect infectious disease threats.² Additionally, search engines such as Google aggregate search queries to predict flu outbreaks³ and Dengue.⁴ The purpose of this project is to identify Internet websites that disseminate near real time information regarding the outbreak of infectious diseases such as influenza (flu), Ebola, severe acute respiratory syndrome (SARS), cholera, West Nile virus, malaria, Escherichia Coli (E Coli) and measles.

Methods
A retrospective review of peer-reviewed published literature and the Internet using the search terms “sentinel surveillance”, “internet”, “emerging/and infectious disease/s”, “syndromic surveillance”, “real-time disease surveillance”, “infectious disease and social media”, and “crowdsourcing epidemiology”. MeSH terms were used for PubMed searches. Articles and websites were scanned for social media tools that support epidemic intelligence.

Results
These websites were identified as among those providing near real time infectious disease information:
Crowdbreak http://www.crowdbreaks.com/
Dengue Trends http://www.google.org/denguetrends/
Flu near you https://flunearyou.org/home
Flu Trend https://www.google.org/flutrends/us/#US
HealthMap http://www.healthmap.org/en/
ProMED-mail http://promedmail.org/

Conclusion
Due to the large number of search returns, pertinent websites were more easily identified by manual review of peer reviewed literature and the Internet websites. The implications from this study suggests a centralized repository is needed for easy access to the information to reduce or prevent the global spread of infectious diseases. Future research could focus on integration and visualization of event-based reporting found in social media tools.

References
A Literature Review of Medication-Related Clinical Decision Support

Clare L. Brown, MPharm¹,², Sarah P. Slight, PhD¹,²,³, Andrew K. Husband, PhD¹, Neil Watson, MBA MSc², David W. Bates, MD MSc³,⁴,⁵

¹Durham University, Stockton on Tees, UK, ²Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK, ³Brigham and Women’s Hospital, Boston, MA, USA, ⁴Harvard Medical School, Boston, MA, USA, ⁵Harvard School of Public Health, Boston, MA, USA.

Abstract
Medication-related clinical decision support (CDS) delivers automated guidance and support to clinicians. We reviewed the literature of basic medication-related CDS functionalities, from 2007 to 2014, using a systematic approach and reflected upon the issues pertinent to future development. Advancements, such as improving alert specificity and application of human factors principles during the design and implementation of CDS, are important considerations to improve patient care and reduce alert fatigue.

Introduction
Medication-related CDS delivers automated guidance and support to clinicians at the point of prescribing, through the use of electronic alerts or order sets and templates. Basic CDS represents functionality, which is typically a starting point for healthcare organisations, examples include drug-drug interaction (DDI) checks, drug allergy checks, dosing guidance, duplicate therapy checks and formulary decision support.¹ CDS has been associated with improved patient safety, improved standards of care and reduced healthcare costs.² We reviewed the recent literature of CDS functionality and reflected upon the issues pertinent to its future development.

Methods
We searched for papers in Medline (Ovid) and Embase (Ovid), using a systematic approach. We included various MeSH terms and key words such as ‘clinical decision support’, ‘electronic prescribing’, ‘computerized physician order entry’ and terms relevant to the five basic CDS functionalities with a date range of 2007 to 2014. We included all publication types, all types of order entry system and all clinical settings. Only English language papers were selected for further review. Titles and abstracts were initially screened, followed by the full text. Reference lists, papers from world leading experts and the ‘other citing articles’ function were also used to identify additional relevant articles. One reviewer performed screening of papers and data extraction.

Results
A total of 896 articles were identified, of which 184 were considered relevant. The success of CDS depends on users finding alerts valuable and acting on the information received. Including more patient-specific parameters to improve alert sensitivity and specificity (i.e. the ability to detect actual errors and the perceived relevance of alerts), and application of human factors design principles is important across all domains. Assigning a severity level to DDI alerts has been shown to improve alert acceptance. Maintenance of accurate records and cross-sensitivity checks are key to the production of appropriate drug-allergy checks. Patient specific parameters should be utilised to improve the relevance of drug-dosage support; furthermore, suggested doses should be appropriately rounded to facilitate administration. How the CDS system is configured is important for the identification of significant drug-duplication errors. The knowledge base(s) for drug-formulary alerts must be accurate and reviewed regularly in order to produce relevant alerts and encourage formulary adherence.

Conclusions
CDS is still undergoing development. The implementation of automation in healthcare has surged in recent years and this is likely to continue. Human-factors design principles and improving alert specificity are important and future research should contribute to this area. Such advancements are important for system developers during the design stage and for end-users of the system who require better functionality and usability to improve patient care and reduce the likelihood of alert fatigue.

References
Developing an Ontology from HIV-associated Elements in Research

William Brown III, DrPH, MA\textsuperscript{1,2}, Chunhua Weng, PhD\textsuperscript{1}, David K. Vawdrey, PhD\textsuperscript{1,3}, Alex Carballo-Diéguez, PhD\textsuperscript{2}, Suzanne Bakken, PhD\textsuperscript{1,4}

\textsuperscript{1}Department of Biomedical Informatics, Columbia University, New York, NY; \textsuperscript{2}HIV Center for Clinical and Behavioral Studies, NY State Psychiatric Institute & Columbia University, New York, NY; \textsuperscript{3}New York-Presbyterian Hospital Value Institute, New York, NY; \textsuperscript{4}School of Nursing, Columbia University, New York, NY

Introduction
In 2012 NIH launched the Big Data to Knowledge (BD2K) initiative to address the lack of appropriate big data tools for translational impact. Two main challenges highlighted by BD2K are “standardizing data and metadata”, and “developing new methods for analyzing & integrating biomedical data”\textsuperscript{1}. Integration of diverse HIV-associated datasets was identified as a research priority\textsuperscript{2}. Integrating biomedical data can increase the breadth of variables and statistical power for analysis; in turn improve knowledge and the effectiveness of interventions. Data harmonization is fundamental to integration, and domain-specific semantic harmonization tools are ideal for managing both the diversity of HIV data elements (DEs) and quickly evolving nature of the HIV research domain. Thus, our aims are to 1) screen HIV DEs and identify common data elements (CDEs), and 2) leverage empirical methods and existing knowledge bases to develop and evaluate an ontology to formally represent CDEs in HIV research.

Methods
We collected the information that describes all data elements in nine HIV-associated studies. Using guidelines from the NIH CDE Resource Portal, we referenced DEs across both empirical and knowledge resources to identify their common data elements (CDEs). Empirical methods included: identification of HIV-associated CDEs by reviewing the research literature, and screening DEs from HIV-associated research data dictionaries, based on NIH criteria. UMLS knowledge resources used included: AIDSinfo HIV/AIDS Glossary and AIDSinfo drug database, as well as CDE from LOINC \(>71,000\) measures, SNOMED \(>311,000\) diseases/measures/medications, and RxNorm [medications]. Protégé was used to construct the ontology.

Results
The HIV-associated Entities in Research Ontology (HERO) comprises 91 HIV-associated CDEs from the research literature, over 850 AIDSinfo HIV/AIDS Glossary terms, and 398 HIV-related drugs from the AIDSinfo drug database. HERO also includes HIV-associated DEs from nine research data dictionaries, as well as formalisms, relationships, and inherited hierarchies from HIV-related CDEs in LOINC, SNOMED, and RxNorm. Results from our final evaluation will include: 1) The amount(s) of different sources that could be mapped to the CDE, 2) The percentage of integrated and non-integrated data from each, and 3) An explication of metadata types.

Conclusion
The next phase includes: identification of overlapping CDEs across knowledge resources, and annotation of multiple CDEs’ source origin(s) and associated formalism(s); concept relationship establishment; and intrinsic and extrinsic evaluation of the ontology for completeness and correctness.

Acknowledgements: This research is supported by NIMH/NIDA (R03-MH103957; Suzanne Bakken, Ph.D.) Dr. William Brown III is supported by NLM research training fellowship (T15-LM007079; George Hripcsak, Ph.D.) and NIMH center grant (P30-MH43520; Robert H. Remien, Ph.D.).

References:
Introduction
Retrospective use of clinical data covers a wide variety of investigations including data exploration, quality assessment, and generation and retrospective validation of hypotheses. While clinicians have the domain knowledge to formulate and test hypotheses, they are limited by the unavailability of user-friendly and intuitive tools giving easy access to collected data. The filtering of datasets to select relevant cohorts of patients is often time-consuming and requires a lot of manual steps and/or advanced informatics skills (e.g. scripting, querying). To streamline the process of filtering datasets, we developed a flexible solution that enables to efficiently and intuitively build cohorts with desired characteristics, perform basic analyses, and export the cohorts to be used in advanced analyses.

The Cohort Selection Application
Our application provides two mechanisms of interaction between the clinical user and the underlying data repositories. Predefined filters can be used for building complex cohorts by clinical users without any need for informatics knowledge. For the power-users (e.g. bioinformaticians) we provide the option to define cohorts by advanced scripting and by directly sending queries to the data repository. Figure 1 shows the two alternatives for building cohorts, by using pre-defined filters and by scripting, for the selection of all triple-negative breast cancer patients in the dataset. By relying on widely-adopted ontologies (e.g. SNOMED-CT) we also incorporate reasoning in the process of building filters and enable users to explore the concepts available in each dataset and their hierarchies, make use of synonyms, etc. Advanced functionality allowing the splitting and merging of defined filter stacks is also implemented. To facilitate collaboration we provide functionality for sharing both the filter-stacks and the corresponding cohorts among teams of clinical users.

Figure 1. Selection of a cohort of triple-negative BC patients by scripting (power-users) and with pre-defined filters
Musculoskeletal Flowsheet Data Modeling for Clinical Research

Matthew D. Byrne, PhD¹, RN; Steven G. Johnson, MS²; Beverly Christie, DNP, RN³; Jung In Park, BS, RN⁴; Lisiane Pruinelli, MSN, RN⁴; Suzan G. Sherman, PhD, RN³; Bonnie L. Westra, PhD, RN, FAAN, FACMI²⁴

¹St. Catherine University, Dept. of Nursing; ²University of Minnesota, Institute for Health Informatics; ³Fairview Health Services; ⁴University of Minnesota, School of Nursing

Abstract: The purpose of the study was to create an ontology representing musculoskeletal concepts using data extracted from a clinical data repository. The ontology connects flowsheet measures to higher-level EHR data concepts to support quality improvement work and research. The structured data derived represents a robust interdisciplinary research and quality improvement opportunity through examination of a large quantity of detailed assessments and interventions that reflect concepts such as mobility, gait, muscle tone, and level of independence.

Objective: Discuss the derivation process and utilization of an ontology representing musculoskeletal concepts using data extracted from a clinical data repository to support quality improvement work and research.

Introduction
Multiple conditions of an acute, traumatic or chronic nature can have an impact on a patient’s musculoskeletal status and functioning. Conditions such as arthritis, which was reported by 22.7% of adults during 2010-2012 (MMWR, 2013), may be exacerbated by overweight/obesity while also having a direct impact on the ability for patients to manage their health and their quality of life. Musculoskeletal conditions may require fine motor assessments of the hand or gross motor functioning demonstrated by a patient’s gait. The timing and effectiveness of assessments of a patient’s ability to be safe and independent are critical to care coordination across the interdisciplinary team. EHR flowsheets vary across vendors, health systems, and service lines, as well as change over time. To compare the health of populations, there is a need to map flowsheet data to higher-level concepts organized by clinical data models (ontologies) to evaluate care and outcomes.

Methods
The musculoskeletal ontology in this study was derived by examining a sample of 66,660 patients with 199,665 encounters that occurred between 10/20/10 and 12/27/13. Data was extracted from a Midwest health system’s EHR and loaded into a clinical data repository (CDR) created by the University of Minnesota. An ontology of musculoskeletal concepts was derived from the data and organized using a functional deficit clinical practice guideline and pertinent clinical questions pertaining to musculoskeletal assessments, interventions and outcomes. An iterative process was used to identify concepts from evidence-based practice guidelines, search for related terms, map flowsheet data to concepts, organize into a hierarchy, and have the process validated by another research team member.

Results/Conclusions
Manual mapping of this complex and overlapping inter-professional dataset resulted in the selection of 783 flowsheet measures. The value sets often had to be reviewed individually to determine the best categorization options. Flowsheet measures found within documentation templates for occupational and physical therapies as well as nursing presented some redundancy and overlap in value sets. Musculoskeletal flowsheet elements within these disciplinary templates also created challenges in that their often discipline-specific assessments had to be matched to and/or differentiated from other musculoskeletal assessment and intervention response sets. The challenges of separating and categorizing some of the concepts by their response sets as well as differentiating the utility of some assessment measures, may be resolved through mapping to standardized terminologies. This poster will present additional challenges and lessons learned in creating a useful clinical data model with related flowsheet data for inclusion in i2b2 and other common data models. The dataset reflects interprofessional care concepts that can be a resource for research, quality assurance/improvement and for quality reporting work like that required by multiple local and federal regulatory and monitoring initiatives. Informatics researchers will benefit by learning from and sharing their experiences in development of clinical data models that support normalization of flowsheet data.

Acknowledgment
Research reported in this poster was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health Award Number UL1TR000114. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Automated searches for personalized evidence to prevent hospital acquired infection

Amos Cahan, MD1,2, Sonya E Shooshan, MLS2, Laritza Rodriguez, MD, PhD2, Dina Demner-Fushman, MD, PhD2

1IBM T.J. Watson Research Center, Yorktown Heights, NY and 2The Lister Hill National Center for Biomedical Communications, National Library of Medicine, US National Institutes of Health, Bethesda, MD.

Objective
Hospital acquired infections result in considerable morbidity and mortality and high treatment costs. Medical caregivers often lack critical knowledge relevant to infection prevention1. We aimed to facilitate knowledge retrieval at point of care by use of an automated search filter accepting terms used in interdisciplinary team notes in an electronic health record system.

Methods
A “Risk for Infection” PubMed® search filter was constructed and its performance compared to an unfiltered search and to two existing search strategies (Table 1) using terms (tokens) extracted from clinical notes in a dataset of de-identified patients from a hospital electronic health record system. Three evaluators independently assessed the returned citations for relevance based on titles alone and titles with abstracts. The fraction of potentially clinically helpful citations of the top 5 citations returned (precision at 5) and inferred average precision were used as measures of search quality.

Table 1: Templates used for PubMed search

<table>
<thead>
<tr>
<th>Filter</th>
<th>Filter Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk for Infection</td>
<td>((Term 1) [Title/Abstract] OR (Term 2) [Title/Abstract] OR (Term 3) [Title/Abstract]) AND infection [MeSH Major Topic] AND &quot;prevention and control&quot; [Subheading] AND Humans [MeSH Heading] AND English [Language].</td>
</tr>
<tr>
<td>InfoBot</td>
<td>Infection AND (Term 1) AND (Term 2 OR Term 3) AND prevention</td>
</tr>
<tr>
<td>Essie</td>
<td>Infection AND ((Term 1) OR (Term 2) OR (Term 3)) AND (prevent^ OR &quot;life style&quot; OR &quot;risk&quot; OR &quot;health behavior&quot; OR &quot;infection control&quot; OR &quot;causality&quot; OR &quot;prevention &amp; control&quot; OR prophylaxis OR prophylactic)</td>
</tr>
</tbody>
</table>

Results
Of 419 notes, 124 had a filled-out Risk for Infection health issue. The Risk for Infection filter provided at least 1 relevant citation of the top 5 results in nearly 2/3 of queries (Table 2). The precision and inferred average precision rates of the filter were significantly higher than an unfiltered PubMed search but lower than a proprietary search engine. Evaluators’ judgment based on title alone reliably predicted judgment by title and abstract (recall 88-99% and precision 88-98% for title compared to title and abstract).

Table 2: Precision at 5 retrieved documents and inferred average precision for 33 queries evaluated using judgments on titles alone and titles with abstracts under strict and relaxed conditions.

<table>
<thead>
<tr>
<th>Filter</th>
<th>PubMed</th>
<th>Essie</th>
<th>Infobot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@5</td>
<td>InfAP</td>
<td>P@5</td>
</tr>
<tr>
<td>Titles</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strict</td>
<td>0.31</td>
<td>0.21</td>
<td>0.08</td>
</tr>
<tr>
<td>Relaxed</td>
<td>0.60</td>
<td>0.21</td>
<td>0.33</td>
</tr>
<tr>
<td>Titles and abstracts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strict</td>
<td>0.36</td>
<td>0.27</td>
<td>0.12</td>
</tr>
<tr>
<td>Relaxed</td>
<td>0.61</td>
<td>0.27</td>
<td>0.28</td>
</tr>
</tbody>
</table>

P@5, precision based on top 5 returned citations; InfAP, inferred average precision. Scoring rules: “Strict” - at least two evaluators considered it relevant or one considered it relevant and 2 “inconclusive”; “Relaxed” - a citation is deemed relevant based on the judgment of one evaluator, or on it being viewed as “inconclusive” by at least two evaluators.

Conclusion
A publicly available tailored PubMed search filter improves the quality of evidence for assessing and managing a patient’s risk of infection in a timely manner. More means to facilitate focused data retrieval by clinicians are warranted.

CDS, EHR and Pharmacogenomics to Estimate Warfarin Dosing

Pedro J. Caraballo, MD, Joseph Sutton, Padma Rao, Kelly Wix, PharmD, Wayne T. Nicholson, MD, PharmD, Cloann Schultz, MS, Carolyn R. Rohrer Vitek, MS, Robert Bleimeyer, Lois C. Hines, R.Ph., David Blair, MD, Mark Parkulo, MD., Mayo Clinic, USA

Introduction

Warfarin is an anticoagulant drug widely used in clinical practice to prevent thromboembolic events. Individual response to warfarin is influenced by multiple factors including variation in two genes (CYP2C9 and VKORC1). This drug also has a narrow therapeutic range and significant side effects; if the dose is too high it can promote bleeding and if it is too low it can allow thromboembolic events. Pharmacogenomics-based algorithms have been proposed to predict the individual initial therapeutic dose and to minimize the time patients are outside the therapeutic range. We present preliminary results of our attempt to develop and implement a clinical decision support (CDS) intervention (calculator/alert) integrated in the workflow to automatically estimate the initial warfarin dose using data from the electronic health record (EHR). We assess the availability of pertinent structured data and the functionality of two commercially available EHRs.

Methods

We used the algorithm published by Gabe et al. in 2008 to calculate the warfarin dose for the first 3 days. The variables and additional information included in the calculator/alert are in the Figure. Different functionality was available in each EHRs to implement the calculator/alert. On one EHR, we used the integrated expert rule system to generate an alert at the time of a “new” warfarin order (no previous warfarin order and normal PT/INR). The alert displayed the calculated dose and additional information. This alert was implemented in clinical practice. On the other EHR we used additional functionality to allow the provider to change data and recalculate the dose on the calculator/alert. However, this was a new untested functionality and production time was significantly longer.

Results

All the variables needed by the calculator could be captured discretely within the EHR by the CDS tool except for the indication for warfarin treatment and INR target. Clinical approval was difficult due to the outcome of recent randomized clinical trials that showed lack of clinical impact of pharmacogenomics testing and recommendations of current clinical guideline that do not support pharmacogenomics testing for warfarin dosing. In our institution, the number of patients with actionable CYP2C9 and VKORC1 results in the EHR in need of new warfarin treatment was very low. Between 9/18/2014 and 1/31/2015, the alert calculated the initial warfarin dose for four patients: #1 The indication was atrial fibrillation, the calculated dose was 6 mg vs. the provider ordered 3 mg; #2 Carotid dissection, 6 mg vs. 5 mg; #3 Prophylaxis for hip arthroplasty, 7.5 mg vs. 3 mg; #4 Prophylaxis for hip arthroplasty, 7.5 mg vs. 10 mg. In all 4 cases the therapeutic dose correlates with the dose suggested by the calculator. The CDS rule also identified 3 patients with VKORC1 results but without CYP2C9 results.

Conclusions

Warfarin dosing calculator can be developed using current functionality in commercially available EHRs. However, several technical and clinical barriers limit efficient implementation in routine clinical practice: lack of preemptive pharmacogenomics testing, lack of support by clinical experts and clinical practice guidelines, and the need of real-time capture of indication for warfarin treatment and INR target. Integration in the workflow is critical for adoption and interactive functionality (calculator/alert) is necessary to improve accuracy and safety.
A graph data model facilitates analysis of collaboration in an emergency department

Matthew B Carson, PhD¹, Stephanie J Gravenor, MS¹, Young Ji Lee, PhD, RN¹, Denise M Scholtens, PhD¹, Conor N Frailey, PhD¹, and Nicholas D Soulakis, PhD¹

¹Northwestern University, Chicago, IL

Introduction: Healthcare is a complex system of interactions between patients and providers on many levels. Collaboration between providers is an integral part of day-to-day operations in a clinical environment. During an encounter, a variety of personnel perform a different but overlapping set of activities in a cooperative effort to facilitate patient care. While an electronic health record (EHR) captures a patient’s medical history and health status, it does not promote the ascertainment of provider collaboration or workflow. Recently, graph databases have been used to store and analyze complex data sets in many industries. Unlike relational databases in which relationships between entities are hidden and can only be revealed by utilizing foreign keys and computationally intensive table joins, a graph data model allows information to be implicitly connected and thus relationships are easily accessible. This makes queries much faster, especially on large data sets. The lack of a rigid table structure makes graph databases highly flexible. Both nodes and relationships can have attributes, which allows for a rich domain representation. In addition, query visualization can easily reveal non-intuitive or distant relationships. These characteristics make the graph database an ideal framework for analyzing and monitoring collaborative patient care.

Objective: To create a high-throughput analysis framework based on a graph data model for identifying and describing provider collaboration in an emergency department workflow at Northwestern Memorial Hospital.

Methods: First, we mapped the workflow schema of business-as-usual care in the emergency department and identified activities performed at each stage in the care process. Second, we designed a graph data model to represent this process (a simple example is shown in Figure 1). Third, we collected data from the Northwestern Medicine Enterprise Data Warehouse (NM EDW), which uses a relational database management system, for 259,289 emergency department encounters from 2012 through 2014 involving 155,976 patients and approximately 11,676 providers. After cleaning the data, we implemented a Neo4j graph database for the set using our model.

Results: We identified > 75 million activities that occurred during the three-year period with an average of 290 activities per encounter. Approximately 1% of patients had > 10 encounters during that time. To ensure query feasibility, we focused on creating an efficient and intuitive graph model. For instance, defining hyperedges as nodes allowed us to identify which provider performed a specific activity during a certain encounter. Modeling time as a series of nodes and edges instead of a node attribute allowed for localized time and date searches. Modeling the workflow as a doubly linked list ensured traversal of steps in constant time. The poster will describe the conversion from workflow to data model, design lessons learned, and examples illustrating how collaboration can be quantified.

Conclusions: A graph data model can be used to analyze collaborative interactions in a complex healthcare environment. By incorporating a graph database, EDWs could provide meta-analyses and monitoring of collaborative relationships and workflows. The model could be expanded to include auxiliary health information such as genomic and proteomic data, regulatory pathways, and precision medicine information for more in-depth analyses.

---

Figure 1. An example of the graph data model showing the activities of three providers during one encounter. Node colors: ED workflow=black, encounter=orange, hyperedge=yellow, providers=green, time=red, activities=blue.
Survey of Mayo Clinic Trainees’ Knowledge, Attitudes, and Opinions Regarding Clinical Informatics

Tomas Carvajal, MD, Karl A. Poterack, MD
Department of Anesthesiology, Mayo Clinic, Phoenix, Arizona

Introduction

In 2010, the American Board of Medical Specialties (ABMS) approved subspecialty certification in clinical informatics (CI); the first board exam was offered in 2013.¹ There are now almost 800 diplomates in CI, but there is little data on what postgraduate medical trainees know about this new specialty. A previous study of medical students showed that only 28% were aware of postgraduate clinical informatics training.² A survey was undertaken to provide further information regarding the knowledge, attitude and opinions of trainees about CI.

Methodology

All Mayo Clinic residents and fellows at all sites (approximately 1300) were sent an email by the Mayo School of Graduate Medical Education asking them to voluntarily participate in an anonymous survey of their knowledge, attitudes, and opinions about CI. The survey included questions about demographic variables, fellowship career opinions, clinical informatics as a specialty, and basic CI knowledge. A mix of multiple choice and single or multiple answer questions were included. Survey was approved by the Mayo Clinic IRB.

Results

A total of 181 trainees completed the survey (14% response rate), with a mean postgraduate training level of 3.5 years. 50% (76/151) were female, and 56% (84/151) between 28-32 years old. 82% graduated from a North American medical school. 85% (27/181) stated their intention to complete a clinical fellowship, and 63% (108/172) believed that fellowships are necessary to attain clinical competence for successful practice. The decision for fellowship has remained stable in 62% (105/170) since medical school.

Only 24% (38/158) were aware of the availability of CI fellowship training, but 71% (111/158) were able to identify the correct definition of CI. Respondents who were not aware of CI training were almost as likely to correctly define CI. (69% v. 74%) There were a total of 70% (336/481) correct responses about the scope of practice of CI. In a mixture of questions about general knowledge based on a CI curriculum the average correct response rate was 63%.

When queried with regard to gathering information about management of a disease process, 33% replied they would initially use a medical reference source such as Pubmed®, 24% stated they would utilize a general web search tool (eg Google®), and 18% would use a specific web site. Of those who would use a specific website, 78% said they would use UpToDate®. These preferences did not show a significant difference between those who were and were not aware of CI fellowship training.

Conclusion

A sample of residents and fellows of different specialties at the Mayo Clinic were surveyed. The majority was unaware of the availability of a CI fellowship, in line with previously published data. However, there was some awareness of the definitions, scope of practice and core content of CI. Increasing the awareness of the fellowship and providing information on the formal training options is needed in the general medical trainee population.

References

A Comparison of Clinical Decision Support Interventions from Commercial and Internally Developed Electronic Health Records

Burcu Celik¹, Pegah Eghbali Alamdari¹, Karen Bavuso RN, MSN², Eileen Yoshida, BScPhm², Saverio Maviglia, MD MS², Thomas Wetter, PhD¹,², Roberto A. Rocha, MD, PhD³,⁴, Charles Lagor, MD, PhD⁵

¹Institute of Medical Biometry and Informatics, Heidelberg University, Heidelberg, Germany; ²Clinical Informatics, Partners eCare, Partners Healthcare, Boston, MA, USA; ³Brigham and Women’s Hospital, Boston, MA, USA; ⁴Harvard Medical School, Boston, MA, USA; ⁵Dept. of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA

Introduction
Healthcare Information Technology (HIT) vendors are increasingly offering Clinical Decision Support (CDS) to meet the demand from healthcare organizations for improving care and efficiency and reducing costs. Although there is an apparent benefit for healthcare organizations of not having to build and maintain CDS it is not evident to what extent their need for rules can be satisfied by vendor delivered inventories. CDS content often needs to be customized by individual organizations, particularly large academic settings. To better understand vendor CDS offerings we classified and compared the type of CDS content provided by a commercial EHR vendor to the internally developed CDS of Partners HealthCare System (PHS).

Methods
We received an inventory of 357 CDS interventions from an EHR vendor. The list included the intervention name, description, EHR module where available, display text, and intent of CDS. We classified the CDS interventions by type using a published taxonomy [1]. We also compared the vendor inventory to a similarly structured inventory of 656 CDS interventions developed internally at PHS, to determine which vendor CDS interventions could be implemented with minimal or no modifications.

Results
Of the 357 vendor CDS interventions analyzed, 10.4% were classified as “close match,” 21.8% as “approximate match,” 31.7% as “no match but related rules exist,” and 36.1% as “no match or related rule.” The first table shows an example of each category. The second table provides their numbers vs degree of match.

<table>
<thead>
<tr>
<th>Type of Match</th>
<th>Vendor CDS</th>
<th>PHS CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close Match</td>
<td>Recommendation for annual eye exam if patient has diabetes</td>
<td>Adult diabetic overdue / almost due for ophthalmological exam, recommend exam</td>
</tr>
<tr>
<td>Approximate Match</td>
<td>Reminder to check serum potassium at least annually if patient has hypertension</td>
<td>If patient is on Thiazide or on an ACE Inhibitor then check potassium</td>
</tr>
<tr>
<td>No match but related rules exist</td>
<td>Patient with rheumatoid arthritis should not be given Methotrexate tablet 7.5 mg daily</td>
<td>Patient has rheumatoid arthritis but no disease-modifying antirheumatic drugs</td>
</tr>
<tr>
<td>No match or related rule</td>
<td>Notification that patient is on Medicare and overdue for certification</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intervention type</th>
<th>Close Match</th>
<th>Approximate Match</th>
<th>No match but related rules exist</th>
<th>No Match or related rule</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care reminders</td>
<td>21</td>
<td>36</td>
<td>29</td>
<td>59</td>
<td>145</td>
</tr>
<tr>
<td>Patient-specific data displays</td>
<td>1</td>
<td>19</td>
<td>33</td>
<td>25</td>
<td>78</td>
</tr>
<tr>
<td>Incomplete documentation alert</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>Drug-age interaction checking</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Other (18 further categories)</td>
<td>5</td>
<td>13</td>
<td>20</td>
<td>32</td>
<td>70</td>
</tr>
</tbody>
</table>

Discussion
Care reminders were the most frequent category, which is not surprising since they are a common form of CDS. However, the results showed that only 32.2% were candidates for implementation (37 Close and 78 Approximate Matches). The results suggest that larger healthcare systems can use vendor inventories as a starting point, but also need efforts to customize and build additional interventions. Smaller organizations with limited resources might benefit from using vendor provided CDS with minimal modifications.

Reference

1434
HealthAlert: A Real-Time Health Monitoring App for Apple’s HealthKit

Basil Chaballout¹, Ryan J. Shaw, PhD, RN²
¹Duke University Computer Science Department, Durham, NC, USA
²Duke University School of Nursing, Durham, NC, USA

Introduction

Apple’s new “Health” app serves as a repository of health and fitness-related data on the iPhone. The app is used in tandem with HealthKit, a programming interface for developers to use accumulated data with their personally developed health applications. The significance of this lies in Apple’s recent partnership with Epic Systems, a medical records company that powers hospitals serving hundreds of millions of patients. Health has for the first time connected the individually operated mobile device to the provider’s electronic health record.

HealthAlert

Our team developed a mobile-based health monitoring solution, “HealthAlert,” that provides real-time health status notifications to a patient, and their chosen social network of family members, friends, or clinicians. This gives users choice over who receives notifications on their health data. It processes data via Apple’s secure HealthKit application programming interface (API) that connects with approved third-party devices and select EHRs. Using certified health-measuring devices, a patient can input his or her health data into Apple’s Health app in order for HealthAlert to analyze the data and send relevant notifications as a text or voice message. If the health metrics are outside of predefined ranges, an alert is sent to the patient’s selected social network indicating that something may be wrong. The ultimate goal of this application is to provide caretakers, clinicians, and family members with real-time reporting about a patient’s health symptoms to ensure that potential health issues are dealt with promptly.

Development and Testing

Our team consisted of undergraduate computer science students, physicians, and professors. Students were responsible for developing an application that allowed the input of the following information: oxygen saturation, heart rate, body mass, and daily activity. Next, we integrated the application with the HealthKit API. This allowed data from HealthKit approved third-party health monitoring devices, to be directed towards the HealthAlert app after it enters the Apple Health interface. This flow of data through Apple’s Health app provides a consistent and certified stream of data. After data is received from HealthAlert, it is analyzed in an external database hosted via a Bitnami server. This server is responsible for updates, additions, deletions, and reads from the external database. In order for notifications and calls to be sent out, the Twilio API on a separate Heroku server was required. The last step was to test the application. The program, which was built on Apple’s Xcode, was archived into an app file with the intent of ad-hoc usage, allowing for the application to be run on a select few devices for testing purposes.

Challenges

Although successful, the team faced challenges associated with decisions regarding real-time status alerts. This included deciding on how to react to data (i.e. hourly vs. daily) efficiently while not creating alert fatigue or sending out alerts from poor quality data (i.e. alerting upon poorly measured blood pressure data). Other challenges were developing an app that does not read data too frequently and drains the phone battery or use too much cellular data. Many of these challenges represented the velocity aspect of Big Data, and the concern with how to deal with data being inputted at an overwhelming rate. There were also development limitations that included only making the app useable with iOS at this time, dealing with costs associated with Twilio messaging, and HealthKit EHR integration in its infancy. Because of regulations and issues with legal systems, the team had to decide not to make an app that is a substitute for emergency contact, but rather just an adjunctive solution to monitoring wellness.

Conclusion

The process was overall a success and received positive reception by the Duke community. Our team has many future developments for the application on its agenda including a clinical trial and integration with our EHR.

References

Influences, Barriers, and Motivations for Healthy Behaviors Among Pediatric Cancer Patients: A Focus Group Approach

Michelle M. Chau, MPH\textsuperscript{1}, Elena J. Ladas, PhD\textsuperscript{1}, and Lena Mamykina, PhD\textsuperscript{1}

\textsuperscript{1}Columbia University, New York, New York

Introduction: Pediatric cancer survivors are joining a growing number of adolescents who suffer from chronic disease.\textsuperscript{1} With more effective therapies, the overall mortality from pediatric malignancy is decreasing. While this is encouraging, survivors are at higher risk for late-effect chronic medical conditions, such as obesity and cardiovascular disease, due to both the underlying condition and side effects of therapy. Interventions to manage nutrition, physical activity, and lifestyle choices can mitigate the effects of such diseases. Adolescents, in particular, confront a unique set of challenges in the self-management of chronic disease. Rapid social, physical, and emotional developments characteristic of this period impose added obstacles for the management of medical conditions. As such, interventions designed to address the distinctive challenges of adolescents with chronic diseases are needed.

Clinical interventions have attempted to utilize technology to improve the nutrition and physical activity of adolescents at risk for chronic diseases. Behavior monitoring, activity logging, goal setting, and gamification are common features used to engage adolescents in health. Researchers have used SMS, email, websites, and mobile applications to deliver such interventions. More innovative interventions are experimenting with image capture to log foods and nutrition. Many of these interventions are school-based and thus, have an inherent social support aspect. However, few applications engage adolescents with social media. Researchers have used posting and delivered interventions via social media. But, researchers have yet to widely incorporate and evaluate social media based intervention components.

Methodology: A focus group was conducted to gain understanding of the behaviors and barriers to healthy lifestyles experienced by pediatric cancer patients. Participants were recruited from the patient population of the Herbert Irving and Adolescent Cancer Center at Columbia University Medical Center. The group included four adolescents and young adults aged 13 to 25 who are in the transition phase (three months before completion of treatment) or are survivors of cancer (more than one year after completion of treatment). The Columbia University Medical Center Institutional Review Board approved this study.

Results: Several prominent themes regarding health management and behaviors emerged from the focus group including: 1) multi-level determinants of health behaviors; 2) initiation and maintenance of health habits; 3) inspirations for health; and 4) the importance of trust in evaluating information resources. In discussing individual, social, and environmental influences on health practices, participants highlighted how personal factors such as experiencing cancer affected health awareness and the development of nutrition and exercise behaviors. Family, friends, and their communities also played prominent roles in influencing how participants eat, cook, and purchase food. In terms of initiating and maintaining healthy habits, focus group participants stressed how simpler routines were easier to maintain. Group members also shared their initial inspirations for transitioning to healthier lifestyles as well as the motivations that keep them on track. When discussing actions to maintain health, many independently developed strategies to satisfy their appetites and cravings and to make decisions in suboptimal environments. Inspiration for health was often found through the use of social media applications and websites. Several discovered ideas for healthy living through photo sharing and social media platforms such as Instagram and Tumblr. Most participants tried fitness and nutrition applications but found them too complicated, not accurate, or lost motivation to continue using them. On the topic of information resources, participants identified trust as an important feature to look for when using applications. The group emphasized the importance of gleaning advice from trusted sources such as clinicians and verified applications.

Discussion: The focus group results suggest several design implications for adolescent interventions. Frequent participant use of social media to search for motivational messages and healthy living ideas calls for a prominent social media component in interventions. Food, recipe, and photo sharing could be useful motivators for adolescents. Personal and environmental influences on health habits suggest the need to tailor informatics interventions for specific dietary needs and individual circumstances. And lastly, trust emerged as an important requirement for information sources and applications. Our findings suggest a strong need to vet information and feature reputable sources in applications.

Acknowledgements: This research was supported by the T15LM007079 grant from the National Library of Medicine.

References
Predicting Clinical Laboratory Turnaround Time
Alex C. Cheng, MEM1, Marc Beller, PMP1, Joshua C. Denny, MD, MS1
1Vanderbilt University, Nashville, TN

Introduction
One challenge with clinical laboratory tests is the variability in the amount of time between when a test is ordered and when the result is available to the provider and patient. Studies have shown that increased mean turnaround time (TAT) and outliers are correlated with adverse outcomes such as increased length of stay in the emergency department. Knowing the time to result could help providers preemptively schedule follow up or mentally prepare to make a clinical judgment at a given time. Previous studies have successfully made improvements on the laboratory workflow that decreased TAT such as using pneumatic tubes for delivery or implementing computerized order entry. The goal of this study is not to decrease TAT but rather to predict the TAT at the time of provider ordering by modeling the process based on previous throughput data.

Methods
We developed our model on data from a pharmacogenomic screening test at Vanderbilt. The data we pulled for this study came from the laboratory information system between March 2013 and December 2014. Each sample had a specimen identification number that we used to track the progression of that sample through the process. We split our data randomly into development and validation cohorts. First, we calculated the throughput time for each sample based on its first and last event timestamp. For the parameters of the model, we considered four factors: the priority tag of the sample, the length of the queue or number of samples in progress when the sample arrived, the time of day the sample arrived, and the day of the week the sample arrived. In the queue length model, we attempted to fit a linear correlation between the queue length and throughput time. For all other parameters we plotted the distribution of TAT and set the prediction window to the 10th and 90th percentile of that distribution. Performance of each model was determined by two measures: the percentage of times the TAT of samples in the validation set fell within the prediction interval, and the size of the prediction interval as a percentage of the median prediction time.

Results and Discussion
We collected 42044 timestamps and for 3789 specimens. The 10th and 90th percentiles for the overall TAT in the development set were 85 hours and 507 hours respectively, making the prediction range 281% of the median TAT. This prediction range is too large to be clinical useful. The prediction ranges for the priority, hours, and days models performed similarly poorly compared to the overall distribution model. We considered narrowing our range by taking the 25th and 75th percentiles, but doing so would hurt accuracy percentages, which were already below 80%. There was a sudden drop in queue lengths after October 2013 that confirmed an operational observation that grants stabilize will give a better picture of the TAT. For any given sequence of consecutive samples, a longer queue at sample reception resulted in a shorter TAT. The laboratory seemed to allow many samples to build up before completing a large batch at once. Standardizing a queue length that triggers the rapid completion of all of the orders would make queue length at reception could be a good indicator of TAT.

Conclusion
We were unable to produce a model that accurately predicted TAT of genetic testing at Vanderbilt. However, we have created a framework for a simple TAT prediction model in clinical laboratory tests based purely on timestamp and priority data. With some modifications, this model could be applied to other laboratory tests. We will need to back up these quantitative measurements with qualitative observations of workflow. These predictive analytics could be coupled with changes in laboratory procedures to improve coordination of care.

References

This research was supported by the National Library of Medicine (T15LM007450-13) and the National Human Genome Research Institute (U01HG007253)
Predicting The Initial Lapses After Alcohol Detoxification Using mHealth
Ming-Yuan Chih, PhD, MHA, University of Kentucky, Lexington, KY

Introduction
To prevent the relapse, a smartphone app, Addiction-Comprehensive Health Enhancement Support System (A-CHESS), offered ongoing support to alcohol addiction patients after leaving the residential care. Previously, some proximal risk factors (e.g., depression, self-reported in the A-CHESS Weekly Check-in) have been shown to predict well the repeated lapses in the coming week. However, the prediction of the initial lapses can be further improved. Based on the literature, demographics, distal risk factors (e.g., quite attempts), and other proximal risk factors (e.g., treatment participation, in this case, A-CHESS use) may predict alcohol lapses. This poster is to report on a predictive modeling study of the initial alcohol lapses among patients using A-CHESS after detoxification. Predicting the upcoming initial lapses can enable A-CHESS to offer instant and preemptive support to the patients.

Methods
Each of the 142 patients recruited from two treatment agencies at the Northeast and Midwest of the U.S.A received a smartphone with A-CHESS and a free 8-month phone/data plan before discharge. Demographics and distal risk factors were collected at the pretest. Proximal risk factors included the recovery progress scores in the A-CHESS Weekly Check-in and the A-CHESS use profiles developed from the A-CHESS use measures. Two datasets—142 initial lapse cases and 95 second initial lapse cases—were prepared for the model development and testing phases. In the model development phase, the Full Model that contained four demographics and 15 risk factors was developed using a logistic regression-based model. Using variable selection procedures, a model that contains only two significant predictors (i.e., gender and recovery progress, called the Simplified Model) was also developed. Model results were compared using area under receiver operating characteristic curve (AUC) analysis.

Results
In the model development phase using the 10-fold cross-validation method, the Simplified Model showed better predictability \((p=0.002)\) with the AUC up to 0.69, as compared to 0.55 in the Full Model (Fig.1). However, the Full Model showed better predictability \((p=0.045)\) using model testing data which were from later lapse episodes, with AUC up to 0.75, as compared to 0.65 in the Simplified Model (Fig.2). Detailed results will be provided in the poster.

**Fig. 1. AUCs in Model Development**

**Fig. 2. AUCs in Model Testing**

Conclusion
The Simplified Model may be better applied to the very initial lapses after patients leaving the facility, while the Full Model for the prediction of the initial lapses in the later episodes. The findings may provide an opportunity to improve the predictability of the existing lapse prediction service provided in A-CHESS. However, future research and more data are needed to further test and apply these findings.

Acknowledgement
This study was supported by grant R01 AA017192 from the National Institute on Alcohol Abuse and Alcoholism.

References
Identifying High Risk of Hospitalization Among Long-Term Care Patients Using Conditional Inference Tree

Si-Chi Chin, Ph.D.
1 SEIU Healthcare NW Training Partnership, Seattle, Washington

Background
High users of health care services are a relatively small group of patients who account for a disproportionately amount of health care utilization. The elderly and disabled, who constituted around 25 percent of the Medicaid population, accounted for about 70 percent of Medicaid spending on services in 2003\textsuperscript{1}. Endeavors to reduce healthcare expenditure have included interventions that target individuals with long-term care (LTC) needs who are at particular risk for hospitalization.

Objective
The study aims to identify a population(s) of high risk of hospitalization. Specifically, this study aims to discover unexpected patient characteristics associated with high risk of hospitalization. The outcome of the study will inform the development of interventions strategies and to customize the interventions to meet specific needs of LTC patients.

Method
Conditional inference trees\textsuperscript{2} have been implemented in this study using data captured from Comprehensive Assessment Reporting Evaluation (CARE) tool. The data contains 76,873 LTC patients that reside in the State of Washington as of 2013. Conditional inference tree is an automated learning algorithm that uses Chi-square test statistics ($p$-value $\leq 0.05$) to measure the strength of association between any given split of a numeric variable (e.g. age or BMI) and the predicted variable (i.e. hospital admissions). The method discovers a sub-group of patients with statistically significant higher risks of hospitalization.

Results
Figure 1 provides two examples of conditional inference tree learning. The visualization shows that patients of age younger than 3 year old are in extraordinary risk of hospitalization (60\% as shown in Figure 1a). The combination of severe depression (Depression score (PHQ9) $> 16$) and morbid obesity (BMI $> 63$) contributes to extreme high risk of hospitalization ($> 60$% as shown in Figure 1b).

![Conditional Inference Tree Examples](image)

(a) Sub-group by age
(b) Sub-group by BMI and Depression Score

Figure 1: Example of conditional inference tree discovery

Conclusion
Conditional inference tree method is shown to be effective in identifying LTC patients with significant higher risk of hospitalization. The analysis suggests additional support (e.g. specialized/skilled care giver, education resource) is high on demand for young toddlers with disabilities and for LTC patient with morbid obesity and severe depression.

---


Privacy Concerns of Internet Users and Implications for Health Information Technology

Hwayoung Cho, MS, BSN, Rebecca Schnall, PhD, MPH, RN
School of Nursing, Columbia University, New York, NY, USA

Abstract Privacy is an important issue for enabling the use of health information technology. This analysis explored privacy concerns of Internet users. We analyzed survey data collected by the Pew Research Center. Most participants were worried about privacy when using the Internet and reported that it was impossible to be completely anonymous. These findings are of particular relevance to persons who use the Internet for meeting health information needs and highlight the importance of addressing privacy and security protections.

Introduction
Seeking health information online has become a popular method for health information seeking. Approximately 81% of U.S. adults reported having access to the Internet in 2012 and 72% of these users have looked online for health information1. Anonymity and privacy is a crucial consideration before use of health information technology.

Methods
This study used a survey dataset of the Pew Research Center collected between July 11 and 14, 2013. This data was collected initially to describe the use of Internet and mobile technology. Researchers conducted a nationwide survey in the U.S. using phone interviews in English of 1,002 adults >18 years old (502 landline and 500 mobile users). Participant selection was by random digit dial of both landline and cell phone numbers. A phone survey was conducted to examine how Internet users perceive anonymity and privacy while using the Internet. Survey questions focused on Internet use, accessing the Internet with a mobile device, sending/receiving email, and use of social networks. We calculated descriptive statistics on the survey data using SPSS.

Results
Sample: 497 (49.6%) participants were male and 505 (50.4%) were female. Most participants (87%) self-identified as White. 792 (79%) were Internet users, and 533 (53.2%) used Social networking sites. Findings: 59% of Internet users did not think it was possible to use the Internet completely anonymously. 18% had tried to use the Internet in a way that masked their identity from certain people or organizations. The most common way was to clear cookies and browser history (64%) (Figure 1). 41% of participants reported that they had deleted/edited something they posted in the past (41%). 35% did not use a website because it asked for their real name (36%). Some respondents reported past experiences that had resulted in adverse outcomes as a result of online activities, such as email/account taken over without permission (21%), trouble in a relationship with family/friends (13%), harassment (12%), stolen personal information (10%), money lost/reputation damage (6%), and physical danger (4%). 68% of survey respondents perceived that current laws were not good enough in protecting people’s privacy when using the Internet.

Conclusion
Findings from this analysis highlight some of the challenges that people currently experience using the Internet and how it relates to concerns over privacy and security. These are of particular relevance to persons who use the Internet for meeting their health information needs and highlight the importance of addressing privacy and security protections in these areas.

Reference
Appropriateness of Overrides of Age-specific Medication Alerts for Elderly Outpatients

Insook Cho, PhD1,2,3, Diane L Seger, RPh4, Sarah P Slight, MPharm PhD1,5, Karen Caputo Nanji, MPH1, Patricia C Dykes, PhD1, Olivia Dalleur, PhD1,7,8, Mary G Amato, PharmD MPH1,6, David W. Bates, MSc1,3,4

1Brigham and Women’s Hospital, Boston, MA, USA; 2Inha University, South Korea; 3Harvard Medical School, Boston, MA, USA; 4Partners Healthcare Systems, Inc., Wellesley, MA, USA; 5Durham University, Durham, UK; 6MCPHS University, Boston, MA; 7Louvain Drug Research Institute, Universite catholique de Louvain (UCL), Belgium; 8Cliniques universitaires Saint-Luc (UCL), Brussels, Belgium

Abstract: Inappropriate medication prescribing by physicians represents an important preventable cause of morbidity and mortality in the elderly. A well-designed geriatric decision support system can improve prescribing by recommending appropriate doses and drugs for this patient group, but prescribers do not always follow these alert recommendations. The objective of this study was to assess what override rates that could be considered reasonable in outpatient settings.

Introduction: Health-care systems worldwide are dealing with increasingly elderly populations, and this is increasing the numbers of patients with comorbidity who are taking multiple medications. One previous study estimated an incidence rate of 27% for adverse drug events in primary care in the outpatient setting, with most problems occurring at the ordering and monitoring stages of care.1 Computerized provider order entry with clinical decision support has been promoted as one of the most promising interventions for improving medication safety, but prescribers do not always follow the recommendations of the alerts, citing their irrelevance and annoyance with interruptions in their clinical workflow. Careful investigations have revealed that many of these alert overrides are clinically inappropriate, while others are clinical justified. However, it remains unclear what a reasonable override rate should be in older outpatients.

Methods and Results: We obtained data on age-specific alert overrides and the coded reasons for overrides cited by providers at the time of prescribing from outpatient clinics and ambulatory hospital-based practices within a large integrated healthcare system from January 2009 to December 2011. For detailed chart review, a group of six trained clinicians developed the appropriateness criteria with good interrater reliability (κ=0.86). We stratified providers according to override frequency and then obtained samples of 200 overrides from the top-25 providers who overrode the most and 600 from the remaining providers. We measured the rate of total overrides and rates of appropriate overrides by drug class according to the ATC (Anatomical Therapeutic Chemical) classification system. We also examined the prescription types (new vs. renewed prescriptions) and the reasons for overrides according to provider.

In total, 11,235 age-specific alerts were triggered by 827 providers during the study period, of which 79.6% (8,943) were overridden. The override rate by the top-25 providers was 97.6%. The chart review revealed that the overall appropriateness of overrides was 40.3% (308/765); it was 50.8% in the top-25 group and 36.6% in the remainder group. Almost one-quarter of the alerts overridden concerned drugs used to treat anxiety (psycholeptics); the appropriate override rate was 16.2% (top-25, 17.3%; remainder, 15.8%). Other frequent overrides with appropriate override rates included neuroleptics (top-25, 72.4%; remainder, 52.0%), anticholinergics (top-25, 4.0%; remainder, 8.0%), and psycholeptics (top-25, 95.7%; remainder, 56.5%). The proportion of renewal orders was higher in the top-25 group (51.8%) than in the remainder group (30.5%). However, for the override reason of ‘patient has tolerated this drug in the past’, the proportion was higher in the remainder group (63.9%) than in the top-25 group (57.9%).

Conclusion: The appropriate override rate of age-specific alerts was four out of ten. Seven out of ten drugs overridden were psycholeptics, neuroleptics, and anticholinergics. The common characteristics of these medications were a long history of being taken and frequent renewals repeating the overrides, which contributed to the high override rate of the frequent overrides. Interventions are needed to improve alert appropriateness and address providers’ overriding behavior in the outpatient setting.

References

System Architecture of CDC I-SMILE Recommendation Engine

Edward Choi, MS¹, Jina Dcruz, PhD², Sizhe Lin, BS,¹ Aashu Singh, BA¹, Hang Su, BA¹, Kelly Ryder, MPH³, Sridhar R. Papagari Sangareddy, MS², Herman Tolentino, MD², Jimeng Sun, PhD¹

¹Georgia Institute of Technology, Atlanta, GA
²Centers for Disease Control and Prevention, Atlanta, GA
³Association of Schools and Programs of Public Health (ASPPH), Washington, DC

Description

Learners or problem-solvers in applied training programs at CDC solve complex tasks that often require knowledge outside their expertise. Moreover, selecting the learning content that best suits the needs for each research can be difficult, because the knowledge to be learned is not in the researcher’s domain. CDC’s I-SMILE recommendation engine is a system with a goal of facilitating a customized retrieval of learning content for users in need of knowledge outside their domain. The architecture of the system is given in Figure 1.

Figure 1. I-SMILE Recommendation Engine System Architecture

The system retrieves learning content assembled from the Internet. The system has indexed course descriptions of massive open online courses (Udacity and Coursera), descriptions of books from Google Books, and PubMed abstracts from the CDC-based learning content management systems. The collected documents are parsed by using natural language processing (NLP) techniques and indexed in a database.

The system uses a hybrid approach of recommendation, a combination of content-based search and user-based search. Content-based search is a conventional method for document searching, which uses a vector-space model as the core method. For example, if a user enters “health information exchange” as the query, the system will calculate the cosine similarity between the query and the documents to retrieve informative documents that contain those words. User-based search is a variation of collaborative filtering, which makes recommendations without considering document contents. For example, let’s assume that John and Mary, after searching for some query q, clicked documents d₁, d₂, d₃ and d₁, d₃ respectively. Then we can assume that Mary will also be interested to have a look at d₂, and the system will recommend d₂ to Mary and other users who use a query similar to q. In addition to traditional collaborative filtering, which only uses information regarding users’ choice of items by click-through data of all users, our method exploits user profiles provided by users. Users can specify learner background, expertise level, and learning style to customize their profile. Queries entered by the user could be expanded using taxonomy or controlled vocabulary provided by the user to increase system recall. For the informatics domain, we tested use of Medical Subject Headings controlled vocabulary for search and filtering. The system includes a mechanism to rate retrieved contents by the end users.
Design guidelines for effective data visualization of sensor monitoring data

Yong Choi, MPH\textsuperscript{1}, George Demiris, PhD\textsuperscript{1}, Arjmand Samuel\textsuperscript{2}, Danny Huang\textsuperscript{2}

\textsuperscript{1}University of Washington, Seattle, WA, USA; \textsuperscript{2}Microsoft Research, Seattle, WA, USA

Introduction

Unobtrusive monitoring of older adults with the use of smart home sensor systems presents a unique opportunity to improve the health and wellness of older adults. However, customizing deployment to address different environments is a key challenge. Furthermore, once a sensor system is deployed, effective processing of a large sensor data set to make meaningful inferences remains another major challenge. Even when successful algorithms have been developed to analyze the data, older adults as end users are not capitalizing on the value added by the data analysis, if findings are not communicated back to them in an effective and comprehensive manner. The goals of this project were: a) to test a platform for sensor system deployment that allows for standardization across sites; b) to design visualization prototypes that demonstrate sensor data; c) assess older adults’ needs and preferences of various visualization approaches; and d) recommend design guidelines for effective data visualization for sensor monitoring.

Methods

This project includes two phases. In the first phase we used the Microsoft Lab of Things (LoT)\textsuperscript{1} platform to deploy a sensor system within a multi-person residence\textsuperscript{2}. LoT is a software infrastructure that simplifies the interconnection of multiple sensor devices across diverse environments and the development of software applications for processing sensor data\textsuperscript{3}. We deployed various sensors (e.g., door/window sensors, multi-sensor collecting data on motion, temperature, luminosity, and humidity) using the LoT platform. Furthermore, we developed a data visualization application that graphs sensor data within the LoT platform to assess the sensor data visualization needs of older adults and their caregivers. We conducted an online survey with 19 older adults and caregivers to assess their visualization needs and preferences. In the second phase, we engage older adult participants within the design of data visualizations derived from the sensor data through a participatory design approach and plan to develop a set of design recommendations based on older adults’ feedback to be reviewed by experts in the field for face validity.

Results

Our evaluation of the visualization application with older adults 65 years of age and older and caregivers in Phase 1 of the project revealed findings that have clear implication for future sensor data visualization. Participants indicated the importance of knowing the time when the last sensor event happened, the amount of time spent in each room, and a total activity level house-wide. In addition, the caregivers and older adults differed significantly in their perceived usefulness of the system. While older adults indicated that the visualization was hard to understand and questioned the usefulness of the system, caregivers expressed that the system was easy to learn and useful for monitoring older adults health and wellness. Currently we are conducting the second phase of this study that includes participatory design and interview sessions with older adults and visualization experts.

Discussion

The smart home sensor data have potential to be valuable resources for older adults and their caregivers in managing health and wellness. However, there still remains a challenge in increasing acceptance and usage of these technologies among older adults. An informative guideline to develop effective data visualization for sensor monitoring can facilitate the adoption of smart home technologies.

References


Acknowledgments

The authors wish to thank Dr. Thai Le and Mr. Christian Bock for their contributions to data collection and analysis during the first phase of the project.
HomeSHARE: A Distributed Smart Homes Testbed Initiative

Kay Connelly, PhD1, Blaine Reeder, PhD2, Amanda K. Hall, PhD3, Kelly Caine, PhD4, Katie Siek, PhD1, George Demiris, PhD3
1Indiana University, Bloomington, IN; 2University of Colorado, Aurora, CO; 3University of Washington, Seattle, WA; 4Clemson University, Clemson, SC

Introduction
Smart homes have great potential to support independent living for older adults as the global population ages1. However, in-home technology researchers typically: 1) conduct small-scale feasibility studies; 2) recruit participants through convenience sampling; 3) and expend substantial resources to build or customize technologies. While important, these efforts often fail to translate or scale beyond their original settings. A current lack of in-home technology research infrastructure prevents investigators from answering research questions that can generalize to larger populations. The Home-based Smart Health Applications across Research Environments (HomeSHARE) initiative seeks to develop a geographically distributed smart homes testbed that will overcome the challenges of conducting large-scale studies. Founding HomeSHARE sites are located at Clemson University, University of Colorado, Indiana University, and University of Washington (Figure 1).

Methods
The HomeSHARE team has extensive experience in informatics, human-computer interaction, privacy, gerontology, user-centered design, and patient-centered research. Our current goal is to characterize the research needs of investigators in these and other areas including artificial intelligence, augmented cognition, and allied health. Starting in September 2014, we have engaged researchers at research conferences and networked through existing professional contacts to determine discipline-specific research questions, governance requirements, and technical issues. International professional events that we targeted included the Gerontological Society of America (GSA) Annual Conference, AMIA Annual Symposium, and AAAI Spring Symposium. Information was collected using individual interviews, focus groups, and online surveys. For example, at GSA we recruited gerontology researchers for half hour interviews or solicited contact information for phone interviews or online survey completion at a later date. This planning phase will complete in August 2015.

Results
A white paper to inform the development phase of the HomeSHARE initiative is in-progress. Researchers from gerontology, nursing, medicine, informatics, and artificial intelligence collectively agree about the potential utility of a geographically distributed smart homes testbed. Researchers recognized multiple potential benefits from a testbed such as standardization across various environments, access to large data sets, and more diverse study populations. They also identified potential challenges that included additional processes before a study could be implemented and overcoming institutional barriers. While research questions will differ by discipline, the same governance issues apply to all investigators who will participate in the HomeSHARE effort. For example, a health researcher might be interested in how sensor-based measures can be used for health-related decision-making to support independence at home while an artificial intelligence researcher might be interested in questions related to automatic system optimization based on emerging performance factors of networked home sites. Governance issues that apply to all researchers include criteria for participation, shared management responsibilities, and data control and sharing.

Conclusion
Full results of the HomeSHARE planning phase will be presented at the AMIA Annual Symposium 2015.

Acknowledgements
This work is supported by National Science Foundation under Grant Nos. 1405682, 1405723, 1405873, 1405951.

References
Towards Personalized Nutrition

Costanza Conti 1,2, Elena Rossi 1,3, Sara Marceglia 4,5, Vittorio Tauro 3, Federica Rizzi 1,2, Monica Lazzaroni 6, Cristina Barlassina 2, Laura Soldati 2 and Daniele Cusi 2,7

1Kos Genetic srl, Milan, Italy; 2Dipartimento di Scienze della Salute, Università degli Studi di Milano, Milan, Italy; 3Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy; 4Dipartimento di Ingegneria e Architettura, Università degli Studi di Trieste, Trieste, Italy; 5Centro Clinico per la Neurostimolazione, le Neurotecnologie e i Disordini del Movimento, Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, Milan, Italy; 6Patologia Clinica, Fondazione IRCCS Istituto Neurologico C. Besta, Milan, Italy; 7Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Italy

Introduction

The role of nutrition in health management is known since ancient times, but its relevance was boosted in the last decade by epidemiological studies pointing out diet as a key risk factor for a number of diseases. Several parallel studies indicated that the traditional approach, focused on caloric and macronutrients intake, fails in capturing essential features of dietary risk factors and specificities associated to the host’s unique genetic background. Nutrigenomics, which already demonstrated that genetic polymorphisms influence response to diet, can offer a powerful approach to unravel the effects of diet on health. This emerging field of research and its translation into prevention strategies needs new tools enabling the systematic collection of a variety of data.

Methods and results

We developed an integrated platform, called Dietary Monitoring Solution (DMS), to collect phenotypic, genetic and lifestyle information linked to a mHealth application providing personalized dietary indications. Our DMS was designed according to the indications of clinicians, geneticists and nutritionists and it is composed by:

a) a web-based platform, based on a framework dedicated to CRF management;

b) a software, calculating the macro and micronutrients intake from each food/meal;

c) a mHealth application monitoring and guiding the subject in his/her ecological context.

The web-based platform is used by the clinician to create the CRFs containing patient’s data (i.e. personal medical history, physical, laboratory examination, therapies, etc.) and by the nutritionist to collect the patient’s dietary habits, through interviews and food frequency questionnaires. Dietary information, collected through 24h-recall interviews, includes a description of foods consumed during each meal and their macro and micro nutrients composition, computed by a software combining input data from dietary interviews and food nutrients composition, from validated food databases. Genotypic data are managed by the healthcare professional and can be uploaded from PED and MAP files, which constitute standard formats for genomic analysis software. The mHealth application is dedicated to the subject and is used to track his/her diet and lifestyle. In addition it provides access to a semi-automatic Decision Support System (DSS), which suggests risk reduction behaviors, considering individual genetic data, particularly those for which there are well established evidences of interaction with the diet. All suggestions provided by the DSS need to be validated by healthcare professionals. This platform was used to collect information from more than 500 volunteers participating to ATHENA, a European Commission FP7 project (Grant Agreement 245121), in three recruitment sites. It also stores volunteers’ genetic data, as generated by genotyping the DNAs with a commercially available chip containing ~280K highly informative genome-wide tag Single Nucleotide Polymorphisms. The user with “Data Manager” role can extract anonymous data from the platform and analyse data, homogeneously collected through the web-based platform, from all the centres taking part into the project.

Conclusion

DMS novelties rely on the possibility for the professional to refine data collection tools, to “supervise” the suggestions proposed by the DSS and to interact with the patients, responding in this way to the concerns raised by healthcare professionals toward the DSS approach. Future studies are envisaged to evaluate users’ compliance and usefulness of the mHealth application suggestions.
User-Centered Design of the Clinical Dashboard for the MySafeCare Patient Safety Reporting System

Brittany Couture1, Jessica Cleveland4, Awatef Ergai PhD4, Zachary Katsulis4, Naoaki Ichihara MD, MPH, PhD2, Ann DeBord Smith MD, MPH1,3, Esteban Gershanik MD, MPH, MMSc1,3, Frank Chang2, Alicia Goodwin5, Sarah A Collins RN, Ph.D.1,2,3

1Brigham and Women’s Hospital, Boston, MA; 2Clinical Informatics, Partners eCare, Partners Healthcare Systems, Boston, MA; 3Harvard Medical School, Boston, MA; 4Healthcare Systems Engineering Institute, Northeastern University, 5OpenClinica, LLC, Waltham, MA, 6Center for Surgery and Public Health, BWH, Boston, MA

Keywords: Patient Centered Care, Patient Safety, User-centered design

Introduction/Background

Patients are a unique source of information on potential threats to safety and quality of health care.1 Past research on patient-perceived safety threats relied on retrospective surveys and interviews. Innovative technology may aid in overcoming barriers that prevent identification of safety threats experienced by patients.2,3 Consumer safety reporting systems exist.4,5 Yet, we know of no system that captures in real time patient perceived threats to safety while in the hospital as a learning health system to mitigate risks before they occur. We are developing MySafeCare (MSC), a web-based/mobile enabled application that provides patients, families and their friends an electronic, real-time way to communicate safety concerns to hospital staff. MSC includes a clinical dashboard for staff to view and trend patient/family reported concerns and document/communicate follow-up. This poster describes our user-centered design for version 1 of the clinical dashboard.

Methods

The targeted users of the version 1 dashboard are Nurse/Medical Directors at Brigham and Women’s Hospital. We conducted 3 Interviews with Nurse Directors and 2 Interviews with Medical Directors. Each interview lasted 30-60 minutes and focused on probing questions to determine the best ways to accurately and efficiently display submitted concerns to dashboard users. The research team analyzed interview notes to extract design requirements by turning them into structured user stories that described functions directors wished to see. An example user story is: “As a dashboard user, I want each case displayed as a separate row on the dashboard, so that I can address individual concerns instead of all of them as a summary”. Next, validated user stories were categorized into minimally releasable features (MRFs), documented using Jira Software (c), prioritized, and confirmed with unit directors.

Results

We defined 25 MRFs after 5 interviews and selected 12 of these to appear in v1. Included in those 12 are: landing page view, status filter, dashboard notifications, and dashboard trending. Each MRF had 1-7 associated specific requirements, for example the drill down case view (Figure 1) includes identification of submitter, concern type, and level of concern.

Discussion/Conclusion

Interviewing targeted users to obtain requirements for dashboard functionality gave us valuable insight to meet our users’ needs. From this qualitative process, we were able to prioritize development efforts on requirements that surfaced during multiple interview sessions, such as trending number or types of concerns in a graphical manner. Through collaboration between our technical team and interviewing team we targeted features that are important to users while also technically feasible to complete before v1 release. This is part of an iterative design process, including design evaluation of each release version to optimize the functionality of the dashboard. The same user centered design methods are being conducted with patients and families for development of a user friendly Patient Facing Application.

Acknowledgements: This work was funded by AHRQ 1P30HS023533-5 Making Acute Care More Patient Centered

References

Design and Evaluation of an Infection-Risk Monitoring Application

Matthew C. Cowperthwaite, PhD, Anurekha Ramakrishnan, MS, Mark G. Burnett, MD, K. Michael Webb, MD, Douglas J. Fox, MD

NeuroTexas Institute Research Foundation, St. David’s Healthcare, Austin, TX 78705

The massive quantities of data generated during routine clinical practice and captured by hospital EMRs hold the great promise to improve the quality and efficiency of healthcare, while also reducing the cost of healthcare delivery. NeuroTexas Institute Research Foundation (NTIRF) at St. David’s Medical Center has been developing and deploying an innovative research database – ProSpect – to capture a wide range of clinical and outcomes data describing neurological and neurosurgical treatments. ProSpect will leverage open-source APIs for rapid implementation of ‘apps’ that support both patients and clinicians.

Currently, ProSpect aggregates clinical data from 6 different hospitals owned by St. David’s HealthCare in Austin, TX, a large neurosurgery practice, and local outpatient imaging providers. The ProSpect project leverages a custom, in-house developed database platform written in SQL. The primary database implements a “schemaless” Entity-Attribute-Value (EAV) model in which data are stored as key-value relationships for each patient. ProSpect collects clinical data including nursing reports, surgery reports, medication reports, CPT and ICD9 codes from St. David’s Electronic Medical Records (EMR). ProSpect also includes a dedicated Merge eFilm (Chicago, IL) Picture Archiving and Communication System (PACS) server for collecting radiological images. Data feeds from various sources are executed and loaded into ProSpect on a daily basis.

Currently, ProSpect contains data for 3,651 neuro-patients treated since 2012. The most common diagnoses are lumbar and cervical spinal spondylosis, displacement of lumbar disc, and spinal stenosis. There are 4,162 distinct neurosurgeries performed on 3,184 patients. 525 patients had a second hospital encounter within a year of surgery, with a median readmission time of 50.5 days. Readmissions may occur for a planned procedure (e.g battery replacement for a spinal simulator) or an unplanned procedure (e.g. a revision or complication).

Our team is interested in surgical-site infections (SSI), and here we present a pilot study that focuses on identification and analysis of risk factors for SSI. Through a literature review and consultation with neurosurgeons and other clinical experts, we identified 50 potential risk factors associated with SSI. The factors were grouped into demographic factors (age, ethnicity etc.), patient co-morbidities (diabetes, hypertension, etc.), surgical factors (surgery duration, blood loss, etc.), and clinical factors (e.g. anemia, inflammatory signals). Entity mapping was done using associated rules-based tagging. The Infection rate was 4.1% (33 out of 790 Lumbar fusion surgeries in 2014 and 1st quarter of 2015). Feature selection was done using Logistic Regression with backward elimination. Class imbalance was handled using Weighted Logistic Regression (WLR) and Synthetic Minority Oversampling Technique (SMOTE). WLR, Decision tree and random forest models were built for classification; with Random forests providing the best precision (positive predictive value) of 70.2% for the infection class. The significant predictors were post-op hemoglobin count, re-fusion surgery, removal of implanted hardware, number of prior surgeries for the patient, Arthrodesis status, smoking and BMI.

The “Infection Risk” app will present a patient and clinician-friendly view of infection-risk based on patient risk factors identified by the model. A prototype of this app is designed using Twitter’s bootstrap, d3.js and JQuery, and runs on standard web, tablet, and mobile devices.

In summary, NTIRF has developed a clinical research database – ProSpect – that gathers and assimilates clinical data spanning the continuum of care to create a detailed portrait of the care delivered and patient recovery and health. We have developed an “Infection Risk” app that will present patient and clinician-friendly view of infection-risk, using ProSpect as the primary data source. We plan on integrating this app to use the open-source SMART API platform in the future.
To Improve Sensitivity and Specificity in Early Detection of Sepsis

Vira S. Danak, MS1, Albert Riedl, MS1, Hien H. Nguyen, MD1
1UC Davis Health System, Sacramento, California, USA

Introduction
The U.S. National Center for Health Statistics of the Centers for Disease Control and Prevention (CDC) has issued estimates derived from claims data that sepsis cases have increased in the U.S. from 621,000 in the year 2000 to 1,141,000 in 2008. With increased use of electronic health records (EHR), a significant amount of secondary EHR data is collected. This data can be used to develop and implement an Early Warning Scoring system (EWS) to identify septic patients earlier and help improve in hospital mortality of septic patients. Since early detection of sepsis in adults is challenging, this thesis explores and proposes opportunities to improve the accuracy of the early warning scoring systems and alerts used to detect sepsis. This study approach is to identify new vital sign thresholds in SIRS criteria for different age groups that can help improve the efficiency of the current sepsis alert at UC Davis Health System (UCDHS). Evaluating these new scoring systems shows improved sensitivity and specificity in early detection of sepsis in adults.

Methodology and Evaluation
The principal objective of the study is to improve the sensitivity and specificity of the current sepsis alert at UCDHS. We use traditional systemic inflammatory response syndrome (SIRS) criteria to alert bedside clinicians to adult patients with possible sepsis and it has helped us meet our goal of improving sepsis related mortality by 15% from 2011-2013; but the non-specific nature of SIRS criteria has led to alert fatigue. In order to improve identification of patients with SIRS and infection, we collected data on patients with clinical infection by asking physicians to answer a question regarding the presence of suspected infection for every patient who is admitted to the hospital from the emergency department. The Infection Answer is basically a categorical variable with Yes/No response. Since a clinician’s interpretation of patient’s condition and judgment is most crucial, this study considers Infection Answer (or Order_Answer) as a gold standard to compare with and evaluate the accuracy of the sepsis alert. To meet this objective and propose future goals for improvement, current sensitivity and specificity of the alert is evaluated/calculated using Structured Query Language (MySQL workbench 6.3) and Statistical Analytical Software (SAS 9.3). Having obtained the current accuracy of the sepsis alert then, data analysis is carried out to explore and meet the other research objectives.

1. The evaluation of current sensitivity of the sepsis alert at UCDHS resulted in 68.13% and specificity of the sepsis alert at UCDHS resulted in 52.46%. Secondly, the alerts results show that alert is over firing based on the findings that around 48% of the order response to the alert is “overridden” by the clinician and the Infection Answer is “No” for those alerts.
2. In this research, pulse is observed to be the foremost participant influencing over firing of the alert and therefore mean pulse thresholds were used to divide the overall adult population in different age groups.
3. Furthermore, results from logistic regression showed that temperature and WBC are associated in determining the infection answer by the clinician. To reaffirm this, the analyses were computed two ways. 1. Independent vital signs participating most frequently in firing of the alert; 2. Vital signs associated in modelling the Infection Answer by the clinician. Both the analysis showed similar results that pulse, WBC and temperature are more associated with infection. Therefore, the new pulse and temperature thresholds were computed with CART analysis in SPM. The specificity highly increases in all the age groups when new Pulse and Temperature thresholds are implemented (Age>=18 and <=39(n=282): 97.70% ; Age<=40 and <=59 (n=616): 95.97% ; Age>=60 and <=79 (n=446): 96.78%; Age >=80(n=86): 98.93% ) These findings can help improve early detection of some patients and make the SIRS criteria more specific.

Conclusion
On the basis of the different analyses, it can be concluded that more accurate EWS and alerts can be developed. This can be achieved by using Age-Pulse association to divide adult sepsis population in different age groups and by implementing different temperature thresholds across these different age groups. It may be possible to increase the specificity of the sepsis alert and reduce the over firing of the alert to avoid alert fatigue among clinical professionals.

Feasibility of Converting the Medicare Synthetic Public Use Data Into a Standardized Data Model for Clinical Research Informatics

Mark D. Danese, MHS, PhD\(^1\); Erica A. Voss, MPH\(^2\); Jennifer Duryea, MPH\(^1\); Michelle Gleeson, PhD\(^1\); Ryan Duryea\(^1\); Amy Matcho\(^2\); Donald O’Hara, MS\(^3\); William E. Stephens\(^4\); Adler Perotte, MD, MA\(^5\); Lee Evans\(^6\); Christian Reich, MD, PhD\(^7\)

\(^1\)Outcomes Insights, Westlake Village, CA; \(^2\)Janssen Research and Development, Raritan, NJ; \(^3\)ComputerAid, Allentown, PA; \(^4\)Signet Accel, Powell, OH; \(^5\)Columbia University, New York, NY; \(^6\)LTS Computing, West Chester, PA; \(^7\)Astra Zeneca, Waltham, MA

Introduction

A challenge in health care informatics is the availability of actual patient data. Virtually all currently available sources of longitudinal data require a data use agreement and data security plan to ensure patient privacy; therefore, it is challenging to develop tools and share best practices that are effective on “real-world” health data. Recently, the Centers for Medicare and Medicaid Services (CMS) released de-identified synthetic Medicare public use data files (SynPUF) to facilitate the development of data analysis tools. To date, the usability of this data for clinical research informatics purposes has not been reported.

Study Aims:

The purpose of this study was to assess the feasibility of using the SynPUF data for clinical research informatics by developing the specifications for conversion to the Observational Medical Outcomes Partnership (OMOP) Common Data Model version 5.0 (CDMv5). A secondary aim was to compare SynPUF to the actual CMS 5% sample limited data set (LDS) to determine the limitations to applying the specification to actual CMS data.

Methods:

We assembled a multidisciplinary, cross-organizational team and developed open-source algorithm specifications to extract, transform, and load (ETL) the raw text files into CDMv5 using publicly available Observational Health Informatics and Data Science (OHDSI) tools. We used the White Rabbit tool to characterize the raw data, and the Rabbit-In-A-Hat tool to create mapping documentation. Each SynPUF variable was mapped to the appropriate table(s) into CDMv5, including variables about people, visits (including hospitalizations), conditions, procedures, deaths, and drugs. We compared the fields available in SynPUF to those normally available in the CMS 5% sample LDS.

Results:

We were able to map or use 282 of 294 (96%) non-empty data fields in the raw SynPUF data to populate at least one table in CDMv5. Fields that were not mapped included 9 annualized cost fields (identical to more detailed cost information already mapped), 1 length of stay field (admission and discharge date already mapped), and 1 date field (inpatient “claim through” date selected instead of “discharge” date). Medical conditions, procedures, visits, drug records and costs accounted for most of the available data. Overall, the usefulness of the data was high with appropriate values in most fields examined. Limitations of SynPUF compared to the 5% sample LDS included missing enrollment dates, limited data on location of care, lack of charged amounts, no modifiers for procedure claims, no specialized laboratory-reporting fields, no revenue codes, and no durable medical equipment, home healthcare, or hospice care data. The file structure also differed across years within LDS, and also between LDS and SynPUF.

Conclusions:

The SynPUF data from CMS represents a usable and valuable tool for clinical research informatics. The resulting specification will support the development of open source ETL software that will enable researchers to explore the SynPUF data in CDMv5. Researchers will then be able to leverage the OMOP standard vocabulary, take advantage of existing exploratory and analytical tools being developed by OHDSI and others, and develop new tools. In terms of our secondary aim, one key limitation is that, because of the missing variables and different data structures, SynPUF does not facilitate a complete “plug and play” process for other CMS datasets including the 5% sample LDS. As a result, additional work will need to be done to adapt the ETL specification to actual CMS datasets. While the synthetic nature of the data prevents it from being used for clinical research purposes, it appears to be useful for methodological research and for testing tools for manipulating and analyzing health care data.
Information System for Mobile Immersive Learning Environment for Just-in-Time Learning in Public Health

Jina Dcruz, MSW PhD², Kelly Ryder, MSW MPH³, Edward Choi, MS¹, Sizhe Lin¹, Aashu Singh, BE¹, Hang Su, BEngg¹, Sridhar R. Papagari Sangareddy, MS², Herman Tolentino, MD², Jimeng Sun, PhD¹

¹Georgia Institute of Technology, Atlanta, GA
²Centers for Disease Control and Prevention, Atlanta, GA
³Association of Schools and Programs of Public Health, Washington, DC

Abstract

I-SMILE (information system for mobile immersive learning environment) is an interactive, immersive, and context aware social learning environment that is based on mobile and cloud platforms. It delivers learner-centered, just-in-time learning for the public health workforce as they solve problems in a challenging and a rapidly evolving environment. From the perspective of public health workforce development, we present a business case, functional requirements, prototype, and lessons from a formative evaluation study.

Description

The practice of public health is rapidly evolving with increased integration opportunities among key stakeholders in an increasingly complex health system, rapid advances in information technology, and an ever growing disease burden across the globe. A competent and capable workforce [1] and the ability to learn in real-time are necessary to address these challenges to maintain and enhance public health capacity and service delivery functions to improve population health. However, a significant crisis exists among the public health workforce fueled by health system and policy changes. Aside from the challenge of an ageing and inadequately trained existing workforce, public health also faces the additional challenge of diminishing resources that constrain recruitment and hiring of a skilled entry-level workforce. Public health informatics capacity building has emerged as a critical workforce need to strengthen the current public health workforce infrastructure and service delivery [2].

Even with a well-structured applied, competency and problem-based curriculum, public health training programs face the challenge of understanding the rapidly evolving training needs in an increasingly complex public health landscape and responding to them in a timely manner. This radical and rapidly changing context requires equally radical and disruptive approaches to applied training. We propose a new paradigm for applied training that is context-aware, learner-centered, and just-in-time. The new paradigm is based on innovations in technology in the areas of mobile computing, machine learning, and distributed computing.

I-SMILE (information system for mobile immersive learning environment) is a collaborative project between the Georgia Institute of Technology College of Computational Science & Engineering and the Centers for Disease Control and Prevention to design, develop, and evaluate an innovative, just-in-time learning system for applied training in public health informatics. On the basis of existing work regarding learning management systems, learning analytics and connectivism as a framework for learning [3], the project is expected to bring together advances in computing and education technology. Technical features of I-SMILE include, but not limited to identification of key domain concepts from problem statements provided by user, search, retrieval and indexing of relevant learning content, Medical Subject heading (MeSH) keyword mapping to increase search relevancy, responsive user interface, user similarity matching, and problem bank generation. Through its unique features and functions, I-SMILE aims to support the evolution of an agile and capable public health workforce in an increasingly complex information rich environment. Usability testing with current CDC Informatics fellows will be conducted to evaluate the ‘proof of concept’ mobile prototype of I-SMILE.

References

The TISS standard for electronic exchange of information in the private health insurance sector in Brazil

Sergio P. de Oliveira1, José F. R. Costa1,2

1 National School of Public Health, Oswaldo Cruz Foundation, Rio de Janeiro, RJ, Brazil 2 National Supplementary Health Agency

Introduction
In Brazil, the health system consists of a complex network of health care providers, forming a public private mix. The Brazilian Federal Constitution gives the State the duty to promote universal and equal access to programs and health services to all Brazilian citizens, guaranteed by the Unified Health System (SUS), which is public. The constitution also states that health care is free to private initiative, through so-called "private health insurance". In this mode, the beneficiaries pay a pre monthly payment with own resources or from your employer, to the providers of private health plans, for healthcare, in accordance with the parameters established in the contract and the specific regulations. The sector is regulated by the National Health Agency (ANS).

Private sector data plans health of Brazil
Over the past decades there has been an increase in demand for private health plans in Brazil. This move coincided with the epidemiological and demographic transition in the country and deepened with the underfunding of public health and the granting of tax incentives by Government. Adding to this, all companies with 30 or more employees must have their own medical service, provided in most cases by a third party provider. In December 2014, there were 1,425 operators in action and more than 72.2 million insurance beneficiaries in private health plans. The services rendered to beneficiaries, they are by means of more than 117,000 establishments of health.

Standard for Supplementary Health Information Exchange - Standard TISS
As a manner to optimize the sector management, in 2005 the ANS established the Standard for Exchange of Information on Health Insurance (TISS). The structure and the improvement of the TISS are defined by consensus on the Standardization of Information in Health Insurance Committee (COPISS), composed of representatives of the agents operating in the private health insurance sector. The exchange of electronic messages uses the International Classification of Diseases version 10 (ICD-10) and the United Terminology for Supplementary Health (TUSS) to identify events and assistance items. The TUSS has 109,640 terms related to events and assistance items. The TISS is organized into five components: 1) Organizational; 2) Content and structure; 3) Concepts of Representation in Health; 4) Security and Privacy; 5) Communication1. With regard to security and privacy, TISS provides for the adoption of digital certificate with SSL or TLS cryptographic protocol, encrypting at least 128 bits, and authentication HASH MD-5 algorithm. The TISS also defines the content and structure of the mandatory electronic messages, defined using the Extensible Markup Language (XML). The TISS is a mandatory standard for the exchange of electronic information between players in the sector, and in September 2014 it became mandatory that the information exchanged with other agents using the TISS being transferred to the ANS, composing an important database of standardized and individualized records in the context of the future Electronic Health Record project at the national level.

Conclusion
Thus was formed wide range of individualized information of the beneficiaries of private health plans in Brazil, reaching the maximum level of disaggregation. This information will subsidize research, management, evaluation and financial monitoring and assistance of operating private health plans. In addition, standardization and the functional and semantic interoperability of health information promoted by TISS collaborate to forming perspective of the Electronic Health Record at the national level.

Reference
Converting the Foundational Model of Anatomy to OWL 2

Landon T. Detwiler, MS\textsuperscript{1}, Jose L.V. Mejino, PhD\textsuperscript{1}, James F. Brinkley, MD, PhD\textsuperscript{1,2,3}

\textsuperscript{1}Departments of Biological Structure, \textsuperscript{2}Computer Science and Engineering, \textsuperscript{3}Biomedical Informatics and Medical Education, University of Washington, Seattle, WA

In this poster we describe the conversion of the Foundational Model of Anatomy (FMA) and the Ontology of Craniofacial Development and Malformation (OCDM) from the Protégé Frames knowledge representation language to OWL 2 (hereafter referred to as OWL). We describe the conversion goals and methodology. Equally importantly, we note what we did not try and accomplish. This conversion is a first step towards tool compatibility, ontology interoperability, and understanding the gains and compromises of such a translation.

The Foundational Model of Anatomy (FMA) (1) is an ontology that is becoming the reference standard for anatomy. To-date the FMA has been represented in Protégé Frames. Although Frames was the representation of choice when we began the FMA project over 15 years ago, the current preferred representation is OWL. Although the conversion of the FMA into OWL has been attempted several times prior, previous efforts were either incomplete, did not meet FMA requirements, or were simply unavailable to the FMA developers. Hence, we implemented our own conversion utility. We have used this tool to convert the FMA into OWL. We are now using the new, converted, FMA in OWL for all new content development.

Our overall conversion approach is implemented as a configurable Java utility. This approach is potentially generalizable to other Frames-based ontologies, which we have partially verified by using it not only to convert the FMA, but also to convert our Ontology of Craniofacial Development and Malformation (OCDM) (2).

The process begins by first converting Frames classes to OWL classes, assigning IRI’s based on existing ontology IDs (FMAID in the case of the FMA) or via an ID generator, and rdfs:labels based on the existing preferred terms. Default slot and slot value conversion strategies are based on value type. These defaults are modifiable via configuration files. Specific example conversion strategies will be described in the poster.

Our primary goals were 1) to capture all or as much as possible of the information in the Frames version, and 2) to make the FMA available as quickly as possible to ontology developers and users. Thus, in the interest of expediency, not all potential improvements were addressed. Finding errors that were not flagged in the Frames version was not a part of this conversion. Additionally, fixing logical inconsistencies or otherwise optimizing the resulting ontology for reasoning is left as a post-conversion task. We expect that these issues will be resolved over time either by ourselves or by other ontology experts, using our converted copy as the starting point.

Because of the fundamental nature of anatomy, and because of the widespread adoption of FMA content by other ontologies and terminologies, the FMA is already an important factor in organizing biomedical data and knowledge. By converting the FMA to the standard representation promoted by the semantic web we ensure that the FMA can become an even more important contributor to the growing worldwide knowledge network in biomedicine.

The conversion code, as well as converted copies of the FMA and the OCDM are available for download.

Conversion code: http://purl.org/sig/software/frames2owl
FMA in OWL: http://purl.org/sig/ont/fma
OCDM in OWL: http://purl.org/sig/ont/ocdm

Funding: NIH DE24417


Predicting the Factors of Improvement of Health Status of Home Health Care Patients: A Holistic Data Mining Approach

Sanjoy Dey, PhD-C²; Katherine Hauwiller²; Pranjul Yadav, MS²; Michael Steinbach, PhD²; György J. Simon, PhD³; Vipin Kumar, PhD²; Connie W. Delaney, PhD, RN, FAAN, FACMI¹; Bonnie L. Westra, PhD, RN, FAAN, FACMI¹,³

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

Introduction
Home health care (HHC) agencies provide health care services at home for patients who need prolonged cares such as seniors, pediatric clients, post-hospital surveillances and thus, can potentially avoid unnecessary hospitalization. To improve the efficiency, patient’s safety, and timeliness of the needed care, Centers for Medicare & Medicaid Services (CMS) has been collecting and reporting the performance data by HHCs electronically through a standardized tool called Outcome and Assessment Information Set (OASIS) since 1999. Mining significant factors and support system characteristics that can lead to the improvement of the overall health of HHC patients is an important research question. However, most of the existing studies use one variable at a time as the potential outcome of the improvement of the health status, and thus miss any potential relationships present among the multiple outcome variables. The goal of this study is to analyze all such potential outcome variables in together to assess the improvement of the overall health of the HHC patients holistically and thus, it can unravel potential relationships present not only among the outcome variables and but also with their underlying associated factors.

Methods
After IRB approval, analysis was performed on large EHRs collected by 581 CMS-certified HHC agencies between October 1, 2009 and December 31, 2010. The dataset contains records of N=270,634 patients in OASIS data format (as described in a previous study [1]) at two different times: one during admission and one during discharge. Both of the records contain demographic, patient history information, high risk factors, living arrangements and caregiver support, health status (e.g., sensory, integument, respiratory, emotional, cognitive, and behavioral), activities of daily living (ADLs) and instrumental ADL (IADLs) and medication management. All of these variables are treated as input to the model. After several stages of preprocessing, 41 binary outcome variables are created to represent the improvement of health status from admission to discharge from HHC. To enhance the interpretability of the model, we further group the 41 outcome variables into eight higher-level groups by summing up the constituent variables of each group. So, each of the eight outcome variables becomes a count variable representing how many of the constituent variables improve. Then, a two-step framework of multi-level learning framework proposed in [2] is used. First, Poisson regression model is used for each of the eight groups independently. Second, each of the outcome groups is relearned, but the prediction obtained from the other seven variables is included as predictors in addition to other generic predictors. This step will elucidate any potential relationships present among the eight variables into the final model. In each of this step, L₁ regularization is performed to select the most predictive factors simultaneously.

Results/Discussion
Preliminary analysis showed that the predictive models had good predictive power, which are far beyond the random model for each of the eight outcomes as shown in the above Figure. In each of the eight outcomes, significant factors obtained from multi-level method were analysed for clinical interpretation. Most of the variables either have been reported in earlier studies or show interesting novel associations. For example, improvement of depression occurs mostly for the patients who were depressed during admission, but did not receive psychiatric nursing and have no drug dependencies. More such associations between the predictors and outcome variables and relationships among the outcome variables are reported.

Reference
Improving Vaccine-Preventable Disease Reporting through Health Information Exchange

Brian E. Dixon, MPA, PhD1,2,3, P. Joseph Gibson, PhD1,4, Shaun J. Grannis, MD, MS2,5
1IU Richard M. Fairbanks School of Public Health, Indianapolis, IN; 2Regenstrief Institute Center for Biomedical Informatics, Indianapolis, IN; 3Center for Health Information and Communication, Health Services Research & Development Service, Veterans Health Administration, U.S. Department of Veterans Affairs; 4Marion County Public Health Department, Indianapolis, IN; 5Indiana University School of Medicine, Indianapolis, IN

Introduction
Vaccine preventable disease (VPD) outbreaks require immediate, effective response. Although clinical providers are legally obligated to report VPD cases under state law, provider reporting is frequently incomplete, error-prone, and delayed. We seek to explore whether an intervention, aimed at providers and delivered electronically through a health information exchange (HIE), can improve VPD reporting rates as well as how such an intervention can be implemented in an integrated infrastructure that includes heterogeneous electronic health record (EHR) systems.

Methods
We are implementing an intervention designed to pre-populate the official state health department VPD reporting form with patient demographics, lab results, and provider information available from EHR system messages routinely captured in a regional HIE. The pre-populated form will be delivered electronically to providers via fax, an EHR system, or HIE-provided inbox based on clinic-preferred workflow. Prior to deploying the intervention, we gathered baseline reporting information from fax, paper, and electronic reports that constitute a reported case and were submitted by both providers and labs to a local health department. We measured the completeness of key reporting data elements separately for paper, fax, and electronic reports, stratifying by report type. We also calculated reporting rates and examined the results stratified by clinical source, disease and report type.

Population Studied
The Marion County Public Health Department (MCPHD) serves a population of 928,000 individuals living in 396 square miles. More than one-quarter of residents (28%) are African American and 10% are Hispanic or Latino.

Results
We collected 4,135 reports (documents) submitted to public health for 3,556 cases of Hepatitis B. Completeness of data elements varied by report type: data element completeness for lab reports averaged 67.6% with a range from 21.9% to 100% (except ethnicity which is less than 1%), while data element completeness for provider reports averaged 64.9% with a range from 20.8% to 100%. Lab report completeness was higher than corresponding provider report fields for 8 of 15 critical fields. Physicians reported in less than 1% of the cases reported by labs, and all physician-reported cases were also reported by labs. We have collected more than 600 additional reports for other VPD cases, including measles, mumps, chickenpox, and pertussis. Analysis of completeness, timeliness, and reporting rates for these diseases is ongoing. We anticipate the analysis will be completed before fall, making the preliminary results available for the poster along with those for HBV.

Conclusion
The rise of EHR and electronic lab reporting (ELR) reporting capacity among health departments due to meaningful use (MU) may improve assessment of disease incidence and burden. Yet data completeness remains problematic for both lab and provider reports, frequently necessitating calls to health systems to investigate and respond to VPD cases. Health information exchanges may help support more complete capture of information while reducing burden for both clinical and public health organizations.
Patient Centered Medical Home (PCMH) Team Huddle Tool

Nhan V. Do, MD¹, Terry J. Newton, MD¹, Rick L. Barnhill BS², Eric A. Shry MD²
¹Office of the Surgeon General, US Army, Falls Church, VA; ²Madigan Army Medical Center, Tacoma, WA

Abstract
The PCMH care model centers on a team care approach. We developed the PCMH Team Huddle Tool (PTHT) to augment the functionality of our EHR to support the delivery of actionable information in the workflow of PCMH care teams. Preliminary results over a 12 month period post implementation show higher percentage rates in 6 HEDIS quality measures in the Army’s Medical Treatment Facilities (MTF) with PTHT than in MTFs without PTHT.

Introduction
The Army has adopted the Patient Centered Medical Home (PCMH) model to improve access, care management, and health outcomes. Successful implementation of the PCMH model centers on a team care approach, and each morning the Army PCMH team “huddles” to coordinate the care for the medical home patients. Prior attempts to meet the information needs of the PCMH team have been either manual aggregation of patient data from up to 5 separate sources or automation through complex electronic dashboards designed by the enterprise with limited user acceptance. Our objective is to augment our Electronic Health Record (EHR) with the development of a health information technology tool that provides rapid access to actionable information within the PCMH team’s workflow. As observed by Bates and Bitton, current EHRs do not have the full functionalities to support the PCMH model of care¹.

System Description
The PCMH Team Huddle Tool (PTHT) was developed at Madigan Army Medical Center through an agile process with close collaboration with subject matter experts from PCMH teams. Workflow analysis and usability studies were conducted as part of the development process. The tool offers multiple functionalities to coordinate the care of the medical home patients such as identifying patients who are transitioning from the Emergency Department or from a hospitalization, flagging overdue preventive services, redirecting appointments to better match patients to their assigned primary care providers, quick look up of clinical data for decision making, and monitoring of performance measures. From June 2013 to May 2014, the PTHT was deployed to 14 MTFs in the Army’s Western Regional Medical Command with 3 MTFs remaining on the deployment schedule. Using the natural experiment setting from the deployment schedule, we compared 4 Healthcare Effectiveness Data and Information Set (HEDIS) screening measures (screening for breast cancer, colorectal cancer, chlamydia, and LDL in diabetic patients) and two HEDIS performance goals (HgbA1c, LDL) between the MTFs with PTHT and those without PTHT.

Results
Achieving HgbA1c goal of < 9% and LDL goal of < 100 mg/dL in diabetic patients were 8.0% (p < 0.0001) and 4.3% (p=0.02) higher respectively in MTFs with PTHT than MTFs without PTHT. Achieving screening goal for breast cancer, colorectal cancer, chlamydia, and diabetics’ LDL were 9.2% (p <0.0001), 5.1% (p <0.0001), 2.2% (p =0.006), and 5.9% (p<0.0001) higher respectively in MTFs with PTHT than MTFs without PTHT.

Conclusion
A well designed information technology tool to serve up actionable information to support a team based workflow has great potential to advance the PCMH model of care. Our preliminary results showed significant differences in 6 HEDIS quality measures over 12 months of implementation of the PCMH Team Huddle Tool, and we are expanding our evaluation of the tool with additional HEDIS measures and additional facilities through an observational study design.

References
Automatic Phone and Text Message Reminders to Increase Patient Completion of Outpatient Laboratory Testing

Trygve Dolber, MD,1 and David C Kaelber, MD, PhD, MPH1,2

1School of Medicine, 2Center for Clinical Informatics Research and Education, The MetroHealth System, Case Western Reserve University, Cleveland, OH

Abstract
Laboratory evaluation is a significant part of patient care. In many healthcare systems, labs are not collected at the point of care, but rather are ordered in clinic with patient instructions to go to a different location. Therefore, many ordered labs may not be collected. We implemented an automated phone and text message reminders system that increased lab collection over 250%. Phone messages were more effective than text messages.

Background
The MetroHealth System is a large, safety-net, integrated healthcare delivery network including a tertiary care medical center and over 20 satellite clinics in northeast OH, with over 1,000,000 outpatient visit and millions of lab tests ordered annually. Most clinics do not collect labs at the point of care, rather ordering labs and instructing patients to go to central locations for lab collection. In 2013 only 8% of patients with outpatient laboratory tests ordered presented to the lab in the final week before their orders expired (1 month after ordering). We implemented automated phone and text reminders to help with lab collection.

Methods
Using our Epic electronic health record (Epic Systems, Verona WI) we identify patients who have had labs ordered but the specimen is still not collected after 3 weeks. We provide this information to a third-party messaging vendor (TeleVox) which sends automated text and phone reminders to these patients in the beginning of the 4th week. Text messages are sent to those patients with a textable phone. Otherwise a phone message was sent. We randomly selected 100 patients (50 phone and 50 text) each from four different two-week periods between 10/14 and 1/15 to determine effectiveness of the messages sent at the beginning of the 4th week on having the lab collected by the end of that week.

Results
Automated messaging results appears in Figure 1. Both verbal phone (37% average effectiveness) and text (29% average effectiveness) messaging (33% overall average effectiveness) were much more effective than the baseline 8% average effectiveness without messaging (p<.05), but were not statistically different (p=0.16).

Discussion
Most evidence suggests that health care-related text message reminders are equally effective as verbal phone call messages,1 but these findings may not apply to a low-income urban setting in the United States. Mobile phones are widely used in low-income communities and may be more stable than landlines,2 and so text messaging could represent a more direct way to reach many patients.

Conclusion
Automated phone and text messages for ordered labs significantly improve laboratory completion.

References
Evaluating Mobile Information Use: Moving from Access to Interpretation and Application

Glynda J. Doyle, RN, MSN1, Karen Furlong, RN, MN, PhD2, Loretta Secco, RN, BScN, MN, PhD2, Judy Bailey, RN, BScN, MN3, Barbara Jamieson, RN, BScN, MN3, Wayne Hansen, EET, BA, MIDST2 & Robert Pafford, BA, Bed, Med2

1 British Columbia Institute of Technology, BC, Canada; 2University of New Brunswick, NS, Canada; 3Cape Breton University, NS, Canada

Abstract

Integration of mobile information supports in nursing curricula is assumed to promote computer and informatics competencies, however it is unclear whether this education strategy promotes evidence-based practice. Since mobile information practices are poorly understood, researchers from three Canadian nursing programs surveyed students and faculty to determine their use of, and attitudes toward, mobile information supports within classroom, laboratory and clinical settings.

Purpose

This multi-site exploratory study evaluated student and faculty perceptions and use of mobile devices and information software programs called Nursing Central © (NC) ™ and uCentral ™ (UC) from Unbound Medicine. With the introduction of entry-to-practice informatics competencies for Registered Nurses1, nurse educators are seeking strategies to foster development of student competency within the rapidly evolving field of informatics. Researchers evaluated student and faculty use of mobile devices and their how information resources supported learning.

Methodology

Data were collected with two online (Qualtrics) surveys designed to measure faculty and student perspectives and use of electronic information accessed via Unbound Medicine software programs. The sample consisted of students and faculty from three Canadian undergraduate nursing programs at six sites where Unbound Medicine products are used. Quantitative analysis included (a) psychometric assessment and determination of validity and reliability of a new Mobile Information Support Evaluation Tool (MISET) with three subscales: usefulness/helpfulness, information literacy support, and use for evidence-based sources, and (b) descriptive and inferential analysis of MISET data. Researchers explored participants’ descriptions of their use, application, and evaluation of NC/UC as teaching and learning resources. Qualitative content from survey text boxes were analyzed for common themes. Researchers were interested in understanding faculty and students’ use and evaluation of mobile within various teaching and learning settings.

Findings

This poster will present the MISET quantitative findings, comparisons across universities, qualitative themes, and triangulation of qualitative and quantitative results. Two emergent themes relevant for nursing education, learning to use mobile resources and student learning outcomes with the use of mobile resources, will be addressed.

Conclusion

The main recommendation emerging from this study is that nurse educators make a paradigm shift away from using mobile technologies to merely support task-oriented, student information seeking, and focus on mobile information technologies to support application of relevant health information to achieve quality patient care and advance nursing knowledge.

References

Using social media data to analyze patient satisfaction of health care facilities

Katherine Doyon BSN, M.S.1, Qing Zeng-Treitler Ph.D1,2, Rebecca Morris MPH1, Catherine Arnott Smith Ph.D3, Yijun Shao PhD1,2

1University of Utah, 2Salt Lake City Veteran Affairs Hospital, 3University Wisconsin-Madison

Abstract: Patient satisfaction has frequently been used as an indicator of health care quality, often determined through surveys. In this study we analyzed social media containing open-ended survey questions to assess patient satisfaction in clinics, hospitals, and other health care facilities in nine regions across the country to identify themes not surveyed in typical health care questionnaires. Our analysis indicated that the majority of patient’s reviews comment on getting care, personal attributes of staff, and the facility.

Introduction: New health care reforms have placed an emphasis on patient satisfaction, often assessed through surveys. Traditional surveys employ Likert scales and multiple choice responses to assess standardized questions. We believe this limits the amount and type of data captured. Our analysis of patient satisfaction differs in that we analyzed open-ended patient comments captured through social media. We analyzed 771 google + reviews from nine different regions across the country to assess the major themes and quantify the primary area of concern for patients and family members.

Methods: 771 Google+ reviews of medical health care facilities were collected from the following 9 regions of the United States: East North Central, East South Central, Mountain, New England, Middle Atlantic, Pacific, South Atlantic, West North Central, and West South Central. This methodology allowed us to have a national sample without any regional bias. These facilities ranged from large academic centers to single provider offices, and were staffed by a variety of physicians from general practitioners to specialists.

Nvivo: A thematic analysis was conducted to identify factors contributing to patient satisfaction. The initial codes were developed using measures found in the Health Resources and Services Administration (HRSA) Patient Satisfaction Survey. QSR NVivo 10 software (QSR International, 2008) was used for the thematic analysis. In addition to the thematic review, an automated topic modeling analysis was conducted. Prevalence of each node was calculated using matrix coding queries of each parent node in each region.

Word Cloud: A word cloud depicting the prevalence of the words in the google + reviews was constructed by the following formula:

\[
\frac{\sum_w c_t_a(w) \times c_t_b(w)}{\sqrt{\sum_w c_t_a(w)^2} \times \sqrt{\sum_w c_t_b(w)^2}}
\]

Results: Overall 771 reviews were coded for analysis. The majority of reviews contain a comment about the provider (41%) followed by aspects of getting care (37%) such as getting an appointment and time spent in the waiting room. Patients were also concerned about the facility itself (23%) and the outcome of the experience (19%).

Conclusion: Our preliminary results indicate that patients comment on aspects of their health care experience outside the typical health care questionnaire. The analysis of the data identified a number of additional topics that contribute to patient satisfaction, not covered by the HRSA Patient Satisfaction Survey.
A Comparative Analysis of Factors for Predicting Inpatient Length of Stay for Colorectal Cancer Patients

1Prerna Dua, 2Akhila Reddy, 3Pradeep Chowriappa

1Department of Health Informatics and Information Management, 2,3Department of Computer Science
Louisiana Tech University, Ruston LA 71272
*prerna@latech.edu

INTRODUCTION

Colorectal cancer is the third most commonly diagnosed cancer and second leading cause of cancer deaths in the United States. Early diagnosis and treatment accordingly can improve the chances of survivability of the patients. There have been developments in treatment of colorectal cancer, such as greater surgical specialization and wider use of laparoscopic procedures. However, these procedures add to the length of the stay of the patient. The model employed in this research attempts to understand the factors that lead to the increased length of stay.

Survival analysis is a collection of statistical methods that deals with the analysis of time duration to until the event of interest occurs. Cox proportional hazard model (also referred as cox model) is a type of semi parametric model that is considered to be the most applicable model to analyze survival data. The main purpose of this research is to illustrate the use of Cox proportional hazard model and to identify the significant features that influence the length of stay of colon cancer patients in the hospitals.

METHODOLOGY

The framework consists of the following major computational steps. (1) In data pre-processing a) Remove outliers and extreme values b) Handle missing data c) Data transformations (2) In Cox model building, we identify the significant features that influence the length of stay of colon cancer patients.

We analyzed the colon cancer data available from Statewide Planning and Research Cooperative System (SPARCS) database. It contains discharge level detail on patient characteristics, diagnoses, treatments, services, outpatient discharge details in New York State from 2009-2012 years. The dataset contains 31 attributes and 10567 records. The explanatory variables, which are assumed to influence the length of stay of colon cancer patients, are shown in the table 1. The response variable is the time until the occurrence of event or the time between two events in days, months, minutes, and seconds. The response variable should contain continuous and interval level data.

Table 1: Categorization of variables for Cox Model

<table>
<thead>
<tr>
<th>Socio Demographic Factors</th>
<th>Admission and Discharge Factors</th>
<th>Health Condition Factors</th>
<th>Geographical and Financial Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Type of Admission</td>
<td>CDS Procedure Code</td>
<td>Source of Payment 1</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>No. of Discharge</td>
<td>APR Discharge</td>
<td>Source of Payment 2</td>
</tr>
<tr>
<td>Race</td>
<td>Discharge Year</td>
<td>Severity Illness Indicators</td>
<td>Facility 4</td>
</tr>
<tr>
<td>Gender</td>
<td>Admission Day of Week</td>
<td>Risk of Mortality</td>
<td>Service Area</td>
</tr>
<tr>
<td></td>
<td>Discharge Day of Week</td>
<td>Medical Surgical Conditions</td>
<td>Operating Certificate Number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Emergency Department Indicator</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total Charges</td>
<td>Total Costs</td>
</tr>
</tbody>
</table>

EXPERIMENTAL RESULTS

The highest numbers of admissions were for the patients (with colon cancer) who are 70 or older and are lowest for the patients aged 0 to 17 years. There was more number of admissions for white, when compared to African american or Hispanic people. The length of stay was longer for the patients aged above 70 when compared to those aged 0 to 17 years. The patients those aged 30 to 69 are 81% less likely to be discharged earlier from the hospital (HR=0.19, p=0.004). There was a significant increase in the LOS as the increase in the age. There were no significant variations in LOS based on gender, ethnicity, and race.

The highest number of admissions was recorded for the patients who were admitted as elective while the lowest was for the patients admitted as urgent. There were more number of admissions on Tuesday and even most of the patients discharged on Tuesdays. Of all the admissions 55% of the patients were discharged to home, 13% of the patients were transferred to skilled nursing home.

The length of stay was shorter for the patients who had emergency admissions (HR=1.064, P=0.044) when compared to those who had electives and are 0.06% more likely to be discharged earlier from the hospital. The Length of stay was significantly shorter for the patients who were discharged on Thursday, Monday, and Tuesday (HR=1.155, P=0.002; HR=1.140, P=0.003; HR=1.090, P=0.056) when compared to those discharged on Friday and are 15%, 14, and 9% more likely to be discharged earlier from the hospital. There were no significant variations in LOS based on patient disposition, admit day of the week. This was probably because less diagnostic procedures were performed as hospital staffing is reduced on weekends. Also, fewer supervisors are present in the hospitals on weekends. Similar results were found in two other studies that were used as a reference.

The major proportion (82%) of the admissions was for the patients who had undergone surgery and also highest number of admissions was for the patients who do not need immediate medical attention. The LOS was significantly shorter for the patients who required immediate medical attention (HR=1.081,P=0.013) and are 8% more likely to be discharged earlier from the hospital when compared to those who do not need. There were no significant variations in the LOS based on severity of illness, risk of mortality, medical and surgical description.
Initial Approach to Creating an Interactive User Interface Design Tool to Enhance User-Centered Design

Kevin R. Dufendach, MD, FAAP; Kim M. Unertl, PhD, MS; Christoph U. Lehmann, MD, FAAP, FACMI
Vanderbilt University, Nashville, TN

Introduction

Visual interfaces in health information technology are often designed by a small group of experts or modeled after existing interfaces. User-centered and participatory approaches to design have the potential to improve satisfaction and the fit between technology and user needs. Including end-users from the start of the design process requires obtaining input from a large number of individuals, which may be logistically impractical. We present a novel approach to collecting user input into health information technology design.

Methods

We developed an interactive graphical user interface design tool to capture feedback from a large number of end-users quickly and easily (Figure). Using a drag-and-drop web-based application, users are able to select data elements, modify them, and customize a user interface quickly and easily.

The design team initially prepares the interface context and customizable elements using a combination of JSON, CSS, and HTML. Users then select which elements they want displayed, position them appropriately, and specify preferences on how they should be formatted. These preferences are reflected immediately on the screen. If desired, the design team can prepare clinical scenarios to give users direct feedback on how their preferences would look under various conditions. Once users are satisfied with the selected elements, display, and location on the screen, they submit their finished products, allowing the design team to aggregate feedback across users and evaluate potential interface approaches.

Application

The first use case planned for study using this tool is a neonatal intensive care unit physician handoff tool, which we will examine in a multi-site study. Neonatal clinicians will select specific elements and format them to their preferences. For example, after selecting “coagulation labs” to appear, the clinician can then format them either as a stick figure or in a list form. Aggregating these data will allow us to delineate the specific information needs from a large number of users.
Regenstrief ePRO: A Rule-Based Platform for Capturing Targeted Patient-Reported Outcomes

Jon Duke, MD, MS1, Daniel J. Vreeman, DPT, MS2,3, Jeremy Leventhal, MS1, Chen Wen, MS1, Katie Allen, BS1, David Haggstrom, MD, MAS1,2,3

1 Regenstrief Institute, Inc., Indianapolis, Indiana
2 Indiana University School of Medicine, Indianapolis, Indiana
3 Veterans Health Administration, Health Services Research and Development Service, Indianapolis, Indiana

Introduction
Self-reported health assessments have received increased attention over the past decade. These appraisals of behaviors, clinical outcomes, and day-to-day health characteristics are now increasingly incorporated into patient care, quality improvement, and research activities. With advancements in technology, shifts have also occurred in creating electronic tools and data systems for capturing Patient Reported Outcomes (PRO). Next-generation electronic PRO (ePRO) systems should incorporate highly targeted delivery of questions to patients based on demographic, clinical (e.g., disease, medications), and behavioral characteristics. The routine, efficient capture of this information could make a significant contribution to patient care and advance research for many disease states. However, connecting disparate data sources, tailoring questions, and capturing patient responses in a usable fashion each present their own challenge.

Methods
Our team set out to develop a comprehensive ePRO tool that was flexible, scalable, and generalizable. This tool would generate and store PRO measures on an unlimited variety of topics. Additionally, the tool would facilitate downstream delivery of responses to the electronic health records, allowing the patient information to be available at the point of care.

Results
A multi-component platform was developed that includes an expandable bank of questions stored in a common data format, a tool for question selection and ad hoc question creation, and an interface for patient data collection. LOINC® (http://loinc.org), a code system for tests, measurements, and observations, was selected to address the standardization of question definition. The platform included integration with the Regenstrief Rule Authoring and Validation Environment (RAVE) for question selection and ad hoc question creation. OpenMRS® (http://openmrs.org), a web-based, medical record software platform, was leveraged as the question delivery mechanism for data capture from patients. The resulting multi-component platform allows user access on a tablet device via the Internet. At the triggering event of a patient check-in (A/D/T message, appointment) the system checks for questions applicable to this patient that have been queued into the “Question Table.” Questions are added to the queue based on demographics and medical record information. Questions can vary in type (e.g. free text, yes/no, multiple choice, scale) and skip logic is included in the overall design. The questions are delivered through an OpenMRS module and are presented to the patient via the tablet. The tool requests the patient confirm their identity prior to displaying the tailored questions. The tool stores patient responses into a database, including the question status (asked/answered), the question response, associated LOINC code, date created, and applicable identifier for future matching. Patient responses are then propagated to the Indiana Network for Patient Care (INPC), a regional health information exchange, for downstream uses.

Conclusion
By utilizing widely adopted standards and open-source web-based applications, we have created a tool that is generalizable to many technology environments. By developing a highly customizable system with the ability to tailor questions based upon disease state or demographics, this ePRO demonstration has created an infrastructure to enable the collection of critical information that will inform both patient care and research. We have proven the feasibility of building a highly adaptable tool and have now begun usability testing in the clinical environment.
Research Objective: Current methods for researchers to obtain patient consent for the authorized use of their data can be inefficient and resource intensive. For this study we conducted a scan of new approaches for obtaining and maintaining patient consent for the authorized use of individually identifiable data for research. This study is part of a larger initiative to inform the development of data capacity for patient-centered outcomes research (PCOR).

Study Design: Findings were informed by a review of relevant literature, thirteen qualitative discussions with key informants, and an in-person expert meeting with twenty-two participants. Our informants for the qualitative discussions and in-person meeting included researchers at patient-powered research networks (PPRNs), academic researchers, federal representatives, private payers, industry experts, and PCORI representatives.

Principal Findings

- Emerging standards for facilitating the sharing of patient authorization include:
  - The Direct protocol, which is a "push" mechanism for securely exchanging data.
  - OAuth, an open standard for authorization that manages identities and would make the request of the data holder on behalf of the patient to share his/her data with a third party.

- Researchers proposed novel approaches for obtaining consent, leveraging technology. These include:
  - "Modular consent," which requires the patient to first provide general consent to a study and then a second level of consent as related sub-studies arise.
  - "Portable consent," which instantiates the concept of "meaningful consent," whereby the consent process and research purpose are clear to the participant. In order to do this, researchers would use software applications with enhanced user interface features that walk the patient through the consent process to help the participant understand what he/she is consenting to.

- Managing Patient Consent. A third party, such as a PPRN, would create “registries” of patients who are interested in and willing to provide data for research on an ongoing basis. This third party would be: 1) responsible for contacting the prospective participants and obtaining documentation of their consent; 2) responsible for building and maintaining the registries of patients who have consented to use of their data for research purposes; and 3) responsible for sharing the registry with potential data holders.

- Consumer-mediated exchange. A patient would have the ability to download his/her medical data to his/her mobile device and push his/her data to a third party (e.g., research organization). This approach could be supported by BlueButton platforms such as the iBlueButton app.

Conclusions: Advances in technology enable new and innovative ways to manage and share patient consent. Institutional Review Boards could assess these methods to make sure they are appropriate in the context of specific proposed studies. These innovative methods for obtaining and managing patient consent have the potential to facilitate the secure sharing of data for PCOR with patient preferences in mind.

Implications for Policy or Practice: Findings from this qualitative data collection could potentially inform investments in PCOR data infrastructure.

1 NORC at the University of Chicago-Bethesda, MD
2 The George Washington University Milken Institute School of Public Health-Washington, D.C
A Six State Review of Grantees’ Experiences with the State HIE Program

Prashila Dullabh, MD¹, Petry S. Ubri¹, Lauren Hovey, MA¹
¹NORC at the University of Chicago, Bethesda, MD

Research Objective
The State Health Information Exchange (HIE) Program aimed to expand the secure movement of electronic health information. We conducted in-depth case studies to assess grantees’ experiences implementing the program, including progress, changes in strategies, challenges, and lessons learned.

Study Design
We selected states with varied programmatic approaches, state contextual factors, and HIE progress, gauged using key outcome measures in clinical lab exchange, care summary exchange, and e-prescribing. We used e-prescribing data from Surescripts, and survey data from the American Hospital Association (AHA) Health IT Supplement and the National Electronic Health Record Survey (NEHRS). From March-May 2014, we conducted site visits and semi-structured discussions with over 108 HIE stakeholders.

Principal Findings

- Grantees pursued dual-pronged strategies to enabling directed and query-based exchange, providing services like care summary and lab results exchange, public health reporting, and admission, discharge, transfer (ADT) alerts.
- Grantees selected federated or centralized infrastructure to meet local stakeholders’ needs and based on the state’s proliferation of HIE organizations.
- Enablers included policy and regulatory levers associated with the Affordable Care Act and payment reform, state and legislative actions, and stakeholder engagement.
- Challenges and lessons learned:
  - Set tangible, intermediate processes and goals that allow for course corrections.
  - Large health systems, that delivered participants and data in some markets and were competitors in others, influenced the uptake of grantees services.
  - Vendor limitations caused some grantees to shift to a best-of-breed approach.
  - Many grantees looked to ongoing development and adoption of standards in the long-term while building data translation capabilities in-house in the short term for interoperability and data capture and quality issues.
- Many grantees sustainable plans focus on Stage 2 meaningful use exchange requirements and aligning service offerings with payment reform priorities.
- Stakeholders credit the program with raising awareness of HIE; facilitating the breaking down of silos; and establishing foundational elements necessary for exchange. Though it is too early to tell the long-term impact, stakeholders felt HIE’s value would increase over time with support from new care and payment models.

Conclusions
Grantees made progress garnering stakeholder trust and participation, expanding HIE service options, and addressing persistent HIE barriers. Notable lessons learned include building exchange capabilities incrementally, selling the HIE concept to big players, and resolving incompatibilities with HIE vendors.

Implications for Policy, Delivery or Practice
States play important roles in leadership and coordination, particularly convening stakeholders, policy development, and needs assessment. In addition, strong, ongoing support related to standards and interoperability is needed; a provider- or federal-led effort to obtain vendor buy-in for HIE goals may be warranted to achieve interoperability. There is also a need to assess how technical solutions evolve in different markets, develop, and disseminate best practices, and develop governance and oversight requirements. Organizations should also monitor sustainability findings and best practices from existing initiatives leveraging pay-for-performance models.
User Adherence to a Web-Based Application for Youth with mild TBI

Jenna R Dyas, BS, Judith W Dexheimer, PhD, Brad Kurowski, MD, MS, Shari L Wade, PhD, Nanhua Zhang, PhD, Lynn Babcock, MD, MS
Cincinnati Children’s Hospital Medical Center, Cincinnati, OH

Introduction
Mild traumatic brain injury (mTBI) is one of the most common injuries sustained by youth in the US accounting for 500,000 emergency department visits annually. Despite the prevalence, there are limited treatments available to help patients cope with symptoms such as fatigue and impaired concentration. We developed a novel web-based intervention that included guidance, individualized symptom management, and evidence-based cognitive-behavioral coping strategies: Self-Monitoring, Activity-Restriction and Relaxation Treatment (SMART). The objective of this study was to evaluate participant adherence to the SMART program.

Data and Methods
All patients 11 to 17 years of age presenting to the pediatric emergency department (ED) from 10/2013 to 7/2014 who presented within 96 hours of sustaining an mTBI were eligible. Participants were asked to log into SMART daily and record their symptoms using the Post-Concussion Symptom Scale (PCSS). Participants also completed web-based surveys assessing symptoms, using the Health and Behavior Inventory (HBI), at baseline in the ED and at 1, 2 and 4 weeks post-injury. Participants were contacted by a research coordinator via their preferred method, (e.g., text) when they were inactive in SMART for more than a few days. Adherence was measured using number of logins and analyzed using descriptive statistics. Participants were considered “completers” if they completed at least one follow-up assessment. The range of usage was correlated using Spearman’s correlation.

Results and Discussion
We enrolled 21 participants and 13 (62%) completed study procedures. Completers and non-completers did not differ with respect to average age (14.3 (SD1.9) vs 15.3 (SD2.4) years), gender (69 vs 63% male), race (100 vs 75% white) or initial symptom burden (24.2 (SD9.9) vs 25.6 (SD13.0)). For completers, the average number of logins per participant was 7 (range: 3-17). The figure shows frequency of logins. Symptoms decreased by 2.2 points each day and only 2 participants (15%) had symptom scores beyond normative values at 2 weeks post-injury. The trend of login frequency decreased similarly but we found no correlation between adherence and HBI or PCSS scores.

Differences in adherence could not be attributed to demographics or symptom burden. The majority of participants were nearly asymptomatic at 2 weeks post-injury; this may have impacted adherence since these participants no longer needed symptom management. For future trials of SMART, confining study procedures to 2 weeks post-injury and employing innovative strategies may ensure good study adherence. Further research could include automated follow-up reminders and more accessible modes of survey delivery, such as a smart phone application.

ABSTRACT
Centralized cardiac monitoring using portable telemetry devices is an important, but costly approach to monitoring at risk patients. Telemetry is highly resource dependent requiring specialized staff and equipment. This safety net is often ordered inappropriately with delays in discontinuation for both adult and non-ICU patients (Bubb, 2011). This in turn creates delays in care for those who most need it.

At our medical center, the decision was made to develop a CPOE orderset with criteria for telemetry placement as algorithmic support for a nurse-driven discontinuation protocol. Analysis of the people, process and technology issues using the sociotechnical model developed by Sittig, et.al. (2010) helped to identify barriers and opportunities for success. This model was selected for its interactive dimensions that lend to improved patient safety. Variations in ordering of telemetry were opportunities for improvement. As nurse driven discontinuation exploration of the socio aspects are critical.

PURPOSE
The purpose of this project was to facilitate discontinuation of telemetry in the adult inpatient setting by implementing a nursing driven protocol using an eight dimensional sociotechnical model as a guide and to minimize inappropriate ordering of telemetry.

SOCIOTECHNICAL MODEL
- Hardware and Software- Multiple order sets, decision support tools for telemetry order entry and discontinuation replaced a single solution.
- Clinical Content- Develop protocols using AHA criteria in ordersets instead of local preference (Drew et al., 2004).
- Human- Computer Interface- Clinical workflow analysis was conducted and a clinical decision support tool was modified for ordering telemetry.
- People- Identified as a rapid cycle improvement project by administrative leadership. Developed protocols that work with physicians, nurse practitioners, nurses, the leader-manager and culture. Interpersonal factors considered included values, culture, and experience.
- Workflow and Communication- Evaluated all the ways that patients were identified as needing telemetry such as ED page, consults, direct evaluations, and referrals.
- Organizational Policies, Procedures- Adult and Pediatrics policies examined for consistency Telemetry guidelines were created using the AHA criteria for patient selection.
- External Rules, Regulations, and Pressures- Local Nurse scope of practice evaluated.
- System Measurement and Monitoring- Pre-implementation identified patients currently on telemetry inappropriately.

CONCLUSIONS
Criterion based ordering and nurse driven protocols offer opportunity to improve the efficiency and effectiveness of ordered therapies and monitoring. The sociotechnical model was helpful in understanding and planning for the eco-system in which provider order entry and fulfillment exists. The interactive socio and technical dimensions exposed corresponding aspects needed in decision making.

REFERENCE


The Evolution of a Clinical Decision Support Request Form

Shirley Xiang Fei RPh, PharmD1, Sharon Bouyer-Ferullo DNP, RN, MHA, CNOR1, Charles Lagor MD, PhD, CPHIMS1, Saverio Maviglia MD, MSc1,2,3, Eileen Yoshida RPh, MBA1

1Clinical Informatics, Partners eCare, Partners Healthcare Systems, Boston, MA, 2Department of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, 3Harvard Medical School, Boston, MA

Abstract

A large hospital organization transitioned from multiple homegrown systems to one commercial enterprise-wide electronic health record (EHR). A standardized clinical decision support (CDS) request form was developed to capture elements of new and existing requests. The form’s content format was significantly revised and the function was redefined to maintain consistency. Lessons learned include organizing the required elements based on EHR functionality and workflow impact of the request form on different stages of the CDS lifecycle.

Introduction

A large academic hospital organization with seven affiliates decided to transition from multiple home-grown and commercial systems to one commercial EHR. The task to develop a standardized request form that would capture necessary elements of CDS was essential to inventory current state CDS, to migrate current state CDS into the new system, to evaluate new CDS requests and to maintain consistency among CDS of different origins.

Methods

We formed a group of clinical informaticians, knowledge Engineers (KEs), and clinical experts who decided on the necessary data elements for a CDS request form. The initial set of requested data resulted in a seven page Word document with embedded macros. This form was significantly revised several times based upon the feedback from requestors, end-users, application teams, and clinical informatics team members. Revisions were further made based on CDS key element requirements, functional completeness, format and usability. The database where this information was kept also changed during this time. The result for managing requests for CDS interventions become a concise, portable, and useful Excel spreadsheet with the minimally necessary data elements that was easily loaded into an open source content management system (JIRA). This system is the source for requests for CDS new and status updates.

Results

The CDS Request Form evolved from a large multi-purpose Word document into an Excel spreadsheet with more concise data element fields. The spreadsheet includes specific tabs that represent different stages of our CDS lifecycle. We are working toward a requester-facing online form which prompts for all necessary elements of the CDS request without overwhelming the users. The governance involved with the CDS request form request and revisions have a specific process in place set by the CDS Committee.

Conclusion

Our CDS transition process included creating an inventory of current state CDS, the transformation and migration of current state CDS into a new commercial EHR system, as well as tracking and evaluating new CDS requests. Satisfying all these different requirements within the short period of time allocated for the implementation of the new system posed a significant challenge. Another challenge was the relative unfamiliarity with the new system, so requirements remained a moving target. While we are still continuously making improvements to the CDS request process, we have learned many lessons during the evolution of the CDS request form so far. We have learned to organize and frame the required elements based on EHR functionality and workflow; we have also learned the impact of the request form has on different stages of the CDS lifecycle.

References

Challenges for Residents in Following Instruction in Laparoscopic Surgery

Yuanyuan Feng, MS\textsuperscript{1}; Hamid R. Zahiri, MD\textsuperscript{2}; Helena Mentis, PhD\textsuperscript{1}
\textsuperscript{1}University of Maryland Baltimore County, Baltimore, MD; \textsuperscript{2}Anne Arundel Medical Center, Annapolis, MD

Abstract

Operating room-based training plays an important role in accumulating experiences in laparoscopic surgeries. Efficient trainings depend on residents’ understanding of the instructions given by senior surgeons. However, residents’ difficulties in understanding instructions are often overlooked by training system design. In this study, we focus on the challenges residents have in following instructions during the surgery. Our findings can be used to make recommendations for the design of effective surgical training system and surgical support system.

Introduction

Simulation-based training, which allows residents to accumulate surgical skills outside the operating room, has become mainstream in surgical education. However, in a recent survey on residents’ perceptions of the simulators, the authors found that residents had difficulties in transferring the skills gained from simulators into clinical settings. Operating room (OR)-based training is the core place where, with the guidance of senior surgeons, residents become proficient in applying surgical skills in real cases. The motivation of this study is to understand the challenges residents have in understanding and reacting to the instructions during a laparoscopic surgery, in order to make recommendations for the design of effective surgical training systems and surgical support systems.

Methods

This is a qualitative, exploratory study. Individual semi-structured interviews were conducted with three senior general surgery residents, who have primarily performed at least one laparoscopic surgery. The interview focused on the residents’ physical and cognitive challenges in following the instructions in their laparoscopic surgery.

Findings

A wide variety of intraoperative challenges for residents in following senior surgeons’ instructions were identified. We classified them into two subsets – cognitive challenges and physical challenges. Cognitive challenges include attention distribution, visually mapping instructions on the video and impaired judgment. Residents identified that they had difficulties in properly allocating their attention, in order to listen to the instructions. Sometimes, after receiving an instruction, they had to stop their tasks at hand and shift their attention to actively listen to the senior surgeon. Residents identified high cognitive load in transferring verbal instructions into visual targets on the video. They needed to ask for more explanations of the instructions and even waited for senior surgeons to physically point to the monitor to locate the correct target. In addition, residents reported to have less confidence in performing a laparoscopic surgery, because they perceived that the laparoscopic surgical environment was uncontrolled. They felt uncertain about how to manipulate the instrument, which instrument to use, how to dissect without sensation, and how to coordinate their perception and movement. Their uncertainty and lack of confidence intimidated them in performing a procedure, limited their surgical skills, and impaired their judgment. The physical challenges include unfamiliarity of the tasks and physical fatigue. Residents reported that instructions sometimes included unfamiliar tasks, which they had not prepared. These tasks resulted in awkward movements, which inhibited them to follow the instructions. Further, all of the residents reported that fatigue was one of major challenges for them to follow the instructions. The heavy workload, static postures, eye gaze and accommodating the senior surgeons preferences added physical challenges for residents to follow instructions.

Conclusion

We identified a variety of challenges residents reported in following instructions in a laparoscopic surgery. The findings indicate that for residents to efficiently transfer their surgical skills from the simulators to the OR, the design of both simulation training systems and surgical support systems should be further enhanced. For simulation system designers, it highlights the needs to train residents in how to deal with these challenges and how to transfer their skills under these difficulties. For surgical support system designers, it addresses the needs to support effective communication among surgical team members and to create an operating environment that reduce the level of uncertainty and stress for residents.
Implementing automated delivery of evidence-based medication safety information to the point of care

Joseph Finkelstein, MD, PhD¹, Hayden Z Adams², Qinlang Chen¹, Kuo Lin¹, Carol Friedman, PhD¹

¹Columbia University, New York, NY; ²Stony Brook University, Stony Brook, NY

Introduction

PubMed provides open access to a vast repository of indexed abstracts containing the most recent literature about medication safety. However access to this information at the point care is limited due to time constraints and search skills of providers. An automated tool available as a mobile app or EMR function that is easy to use and that rapidly provides highly relevant safety information can address these barriers.

Methods

NLM APIs and MESH indexing was used to develop specialized medication queries addressing major types of medication safety issues. The ultimate goal is to develop and validate a set of clinical filters which support automated retrieval of medication safety information at the point of care.

Results

The prototype interface is presented in Figure 1. Depending on patient history, information can be provided on possible adverse drug reactions (ADRs) as well as various interactions. Preliminary analysis of medication filters to identify adverse drug reactions demonstrated satisfactory recall and precision. Evaluation of one of the filters demonstrated a recall of 90% and precision of 93% in retrieval of all ADRs associated with a specific medication¹.

Figure 1. Automated extraction of medication safety rules from PubMed.

References

Grantees’ Lessons Learned in Implementing State HIE Initiatives

Kathryn Fischer, MPH¹, Prashila Dullabh, MD¹, Lauren Hovey, MA¹,
¹NORC at the University of Chicago, Bethesda, MD;

Research Objective

The State Health Information Exchange Cooperative Agreement Program (State HIE Program) was a four-year program under the Office of the National Coordinator for Health Information Technology (ONC) that provided the 56 states and territories with $564 million dollars in funding to invest in HIE solutions. ONC charged states and territories or state designated entities (grantees) with creating or enabling the necessary governance, policies, technical services, business operations, and financing mechanisms to advance HIE. We analyzed grantee final reports to identify key drivers that helped facilitate and/or accelerate HIE under the State HIE Program as well as key challenges grantees encountered, strategies grantees employed or are considering to overcome those obstacles, and remaining gaps in connectivity.

Study Design

At the conclusion of the State HIE Program in spring 2014, ONC required grantees to submit a close-out report discussing lessons learned, a gaps analysis, and recommendations. Grantee responses were imported into the qualitative research software program NVivo and coded into categories representing the most commonly reported HIE key drivers, challenges, and solutions in the reports.

Principal Findings

Across grantees, the most commonly reported key drivers for successful HIE initiatives were diverse stakeholder engagement to build relationships, establish trust, improve acceptance, align goals, leverage existing HIE assets, and create a sense of ownership for all those involved; strong initial service offerings to drive momentum; clear and consistent policy and regulation; continued marketing of benefits to build awareness and increase adoption; a flexible HIE infrastructure to cope with the evolving nature of health IT; and leveraging payment reform initiatives in order to achieve a critical mass of HIE participants and ensure the utility of HIE services.

The most commonly reported challenges for HIE initiatives were the unexpected resource intensiveness, in terms of time, money, and efforts, for all stakeholders; competing regulations making HIE prioritization difficult; performance limitations of health IT vendors and HIE products; and patient acceptance and lack of common consent policies; and workflow pressures for health care workers adjusting to and incorporating new technology.

Conclusions

The State HIE Program catalyzed progress toward enabling HIE services through a substantial, one-time infusion of funds. However, gaps remain to achieving interoperable intra-state, inter-state, and national HIE exchange. Grantees encountered cross-cutting challenges while implementing HIE activities and reported innovative solutions that can guide and inform similar efforts moving forward and provide an opportunity for grantees to learn from each other’s experiences as they work towards sustainable HIE initiatives.

Implications for Policy, Delivery or Practice

By identifying key drivers, these findings will help policy makers design future programs, service offerings, and policies to advance HIE. Discussing the challenges grantees encountered can help prepare stakeholders for future challenges and set realistic expectations. Grantees can also identify successful strategies used by others to efficiently address their own challenges.
A Baseline Assessment of the Dispensary Workflow in the Birmingham Free Clinic: A Time-Motion Study of Pharmacist Tasks

Arielle M. Fisher¹, Michael Q. Ding¹, Harry Hochheiser, PhD¹,², Gerald P. Douglas, PhD¹
¹Center for Health Informatics for the Underserved, Department of Biomedical Informatics, School of Medicine, ²Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA

Introduction
Healthcare services in low-resource settings often face distinctive informatics challenges arising from health, economic, and education disparities in underserved populations. The Birmingham Free Clinic (BFC), in Pittsburgh, PA is a free, walk-in clinic offering primary care and pharmaceutical services to a medically vulnerable population. The recent implementation of an electronic medical record (EMR) has produced benefits such as improving continuity of care for patients in the University of Pittsburgh Medical Center (UPMC) network. However, pharmacists report that the EMR deployment has had a negative impact on their workflow efficiency. After a qualitative investigation of these negative impacts uncovered 13 workflow inefficiencies, we conducted a time-motion study to provide an evidence-based understanding of the magnitude of these workflow challenges. The results of this study will guide the development of an information system at the dispensary level grounded in several interconnected informatics interventions. Our results will also serve as baseline data for a similar post-implementation time-motion study which will evaluate the impact of the information system on improving pharmacist efficiency.

Methods
Qualitative observations of pharmacist work at the BFC informed the creation of five task categories and 12 subcategories for our time-motion study. Two researchers used the Time Motion Study application for Android devices to capture different pharmacist activities and their specific durations. Two 3-hour pilot sessions were conducted to compare the agreement between researchers prior to starting data collection. Cohen’s kappa was found to be $\kappa=0.808$ following the pilot sessions. Time-motion data was collected in three sessions between October and November 2013; each session lasted approximately three hours. We conducted a brief interview with the pharmacists to identify tasks from our list of subcategories that they believe add direct value to patient care. Based on this assessment, we calculated the value quotient for each dataset in order to quantify inefficiency in the dispensary workflow. The value quotient is a metric for determining a health system’s success in meeting patient needs while effectively using the resources they have been granted¹.

Results
The five workflow categories and their respective time investments were: patient diagnosis (21.5%), EMR-related functions (14.8%), prescription preparation (39.8%), patient interaction (18.7%), and other (5.2%). Our results showed that the pharmacists invested, on average, the majority of their time into prescription preparation which included five subcategories: traveling (2.4%), hunting for medication (4.8%), dispensing medication (7.3%), labeling medication bottles (21.8%), and duplicate documenting (3.6%). Further, the overall value quotient was found to be 43.5% which demonstrates that over half of the pharmacists’ time is spent completing tasks that they consider to not be directly benefiting patient care.

Conclusions
We believe that this is the first study to quantify how pharmacists spend their time in a low-resource dispensary. Our results show that pharmacists at the BFC spend the majority of their time preparing prescriptions, which is a task they consider to be inefficient. The value quotient demonstrates that there are many non-value-added tasks that we believe could potentially be reduced through the introduction of workflow re-engineering.

References
Transforming the National Department of Veterans Affairs Data Warehouse to the OMOP Common Data Model

Fern FitzHenry\textsuperscript{1,2}, Jesse Brannen\textsuperscript{1}, Jason Denton\textsuperscript{1,2}, Jonathan R. Nebeke\textsuperscript{3,4}, Scott L. DuVall\textsuperscript{3,4}, Freneka Minter\textsuperscript{1,2}, Jeffrey Schneit\textsuperscript{3}, Brian Sauer\textsuperscript{3,4}, Lucila Ohno-Machado\textsuperscript{5}, Michael E. Matheny\textsuperscript{1,2}

\textsuperscript{1}Tennessee Valley Healthcare System, Veterans Affairs Medical Center, Nashville, TN; \textsuperscript{2}Vanderbilt University, Nashville, TN; \textsuperscript{3}VA Salt Lake City Health Care System, Salt Lake City, UT; \textsuperscript{4}University of Utah, Salt Lake City, UT; \textsuperscript{5}Bioinformatics and Systems Biology, University of California, San Diego, CA;

This poster describes the conversion of the national Department of Veterans Affairs (VA) healthcare network's corporate data warehouse to a common data model (CDM) suitable for distributed observational research: the Observational Medical Outcomes Partnerships (OMOP) CDM. The CDM allows shared data analysis to have syntactic and semantic interoperability through standardized terminologies such as ICD, RxNorm, SNOMED, CPT, HCPSC, and LOINC. Observational outcomes from electronic medical record systems are becoming more important in comparative effectiveness research, particularly as post marketing surveillance research.\textsuperscript{1} Although randomized clinical trials (RCTs) are the most prominent and strongest research strategy for determining safety and efficacy of new treatments or products, RCTs are expensive, slow, and often do not represent the real world conditions in which the product will be used.\textsuperscript{2} In addition, RCTs may not represent the real world population that will use the treatment or last long enough to identify all the adverse events associated with the treatments.\textsuperscript{2}

The VA has the largest integrated healthcare system in the US with electronic health record coverage from the late 1990s. It is one of the few healthcare systems with extended longitudinal records of covered Veterans and has continuity in patient coverage because eligibility is not related to changes in health insurance seen in other health systems. The VA corporate data warehouse has undergone an initial transformation into OMOP CDM with large-scale data population beginning in 1999 (1998 for drugs). There are approximately 16,927 thousand unique patients in the dataset (see Table 1), with 11,368 thousand of those having at least one encounter (5,559 thousand patients enrolled with no encounters).

Drug cost and procedure cost are pending load. As part of a cooperative effort with the patient-centered Scalable National Network for Effectiveness Research (pSCANNER), we are also transforming the OMOP CDM to the PCORnet data model. VA's OMOP data will be available for initiatives of VA researcher groups and research initiatives of PCORnet. The Observational Health Data Sciences and Informatics (ODHSI) community is developing a suite of data characterization, data quality, cohort generation, comparative effectiveness analysis, and surveillance tools (\texttt{www.ohdsi.org}), and VINCI plans on participating in tool development and supporting use of these tools within the computing infrastructure to support observational cohort analytics. For example, an ODHSI shared quality tool indicates 255,552 records of Sodium Chloride 9mg/ml irrigation solution had a quantity of 1000, e.g. source data used ml vs. bag to record quantity. Strategies to address these types of variations are a future challenge. Financial Support: This study’s work supported with resources and the use of facilities at the TVHS VA and VA SLCHCS and is funded by PCORI contract CDNR-1306-048189 and the VA Informatics and Computing Infrastructure (VINCI) HIR 08-204 and RES 13-457. Prior work described in this paper was supported by AHRQ grant R01HS019913 and NIA grant 1RC4AG039115-01 as part of the American Recovery and Reinvestment Act, National Center for Biomedical Computing (Grant U54 HL108460), and VA HSRR&D CDA-22008-020.


---

<table>
<thead>
<tr>
<th>Table 1: VA Domains and Record Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domains</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Persons</td>
</tr>
<tr>
<td>Deaths</td>
</tr>
<tr>
<td>Organizations</td>
</tr>
<tr>
<td>Locations</td>
</tr>
<tr>
<td>Care Site</td>
</tr>
<tr>
<td>Provider</td>
</tr>
<tr>
<td>Visits</td>
</tr>
<tr>
<td>Procedures</td>
</tr>
<tr>
<td>Drug exposures</td>
</tr>
<tr>
<td>Drug Era</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Observation Period</td>
</tr>
<tr>
<td>Conditions</td>
</tr>
<tr>
<td>Condition Era</td>
</tr>
</tbody>
</table>
Determining Factors Relevant to Handoff by Role and Patient Population

Karol L. Fleming MBA, JD,1 Ray McKelvy,1 Nathan C. Gonzalez,1 William B. Webb BSPH, MBA,1 Robert L Shapiro MD, MSCI1,2

1 MEDarchon, Inc. 113 29th Avenue South, Nashville, TN 37212
2 Cardiovascular Division, Washington University in St. Louis School of Medicine, 660 S. Euclid Avenue, St. Louis, MO 63110

Background: While electronic handoff tools can be significant,1 a review of the literature has found significant variability in the factors that should be included. Many e-handoff tools provide a one-size fits all solution and do not consider the specific information needs of users based on their role or patient population. Methods: Doctors and nurses with at least two years of clinical experience were selectively recruited based upon their role, specialty and patient population. After informed consent, subjects were asked to narrate a clinical handoff. They then were presented with a number of factors identified from the clinical literature as potentially important to handoff and asked to classify these factors as things that should always, sometimes or never be included. Subjects were also asked to describe the circumstances that would lead to including factors listed as “sometimes” and to list any factors not included. These responses were codified and summarized using descriptive statistics and the Z score. The kappa coefficient was used to demonstrate the agreement between whether a participant included a factor in their vignette and how they subsequently categorized that factor (always include vs. never include) in the prompted exercise. Classification and Regression Tree modeling (© Salford Systems) was used to determine whether a factor should be presented based on user attributes, patient attributes, and other variables gleaned from the narrative surrounding when to include variables classified as “sometimes.” Results: 23 of 24 recruited subjects completed the study. Figure 1 shows the percentage of participants who included each factor in their handoff vignette, while Figure 2 demonstrates the extent to which subjects indicated they would include factors. The average kappa for intra-observer reliability was 0.21. Participants from the surgical setting showed poorer concordance between vignette and classification (κ=0.12) than those from the ICU (κ=0.21) or medical settings (κ=0.26). In subgroup analysis, practitioners in the ICU were more likely to include urine output (p=0.06) and less likely to included patient condition (p=0.05); medical physicians were more likely to include a patient picture (p=0.06) and less likely to include the unit admit date (p=0.07) or surgical procedures performed (p=0.03); physicians were more likely to include patient condition (p=0.10) and less likely to include precautions (p=0.01), pain medication information (p=0.04) and urine output (p=0.05); and clinicians seeing pediatric patients were less likely to include patient condition (p=0.03) and room number (p=0.10). Factors suggested by subjects but not found in the literature included information related to: the patient’s family, current care team, catheters, lines and drips, maternal/gestational history, recent events, and discharge planning. Subjects indicated more often than not that those factors categorized as “sometimes” were only discussed when abnormal. Models were successfully produced for each handoff factor found in the literature. Conclusion: This study suggests several new factors that are important for inclusion in an electronic handoff tool as well as the complexity of building such a tool. In order to optimize the user experience and use related hazards, an e-handoff tool must contemplate the user’s role, specialty, patient population and the abnormality of data when deciding whether to display a certain factor.

Figure 1: Factors in Clinical Vignette

Figure 2: Factors in Guided Analysis

Bridging the MedlinePlus Cloud to askMEDLINE

Paul Fontelo, MD, MPH, Fang Liu, MS
National Library of Medicine, Bethesda, MD

Abstract
We developed a Web interface to show the top searched terms from MedlinePlus Cloud and direct the top searched terms to askMEDLINE’s query database. The bridge from MedlinePlus, a patient and family health information resource to askMEDLINE provides recent evidence from PubMed for patients and health care professionals.

Introduction
MedlinePlus is online health resource for patient and families. The MedlinePlus Cloud provides the list of top 100 searched terms from patients1. It is updated daily. The results are displayed with text cloud, a visualization tool to quickly learn the most searched terms in the list. askMEDLINE was designed to search PubMed using free-text queries without knowing complex search strategies2. askMEDLINE has an additional feature that displays the previous user queries. We wanted to build a channel between MedlinePlus top searched terms to the askMEDLINE database.

Methods
The source code from the MedlinePlus Cloud Web page was parsed using a script and then the top 100 terms and their respective rankings were retrieved from the source code. A Web interface was generated to dynamically display by alphabetical order or ranking using text cloud visualization (top ranked terms in the larger font format)3. Users can choose to view the terms either by rank order or in the alphabet. A link was embedded with each term to redirect to previously searched questions in askMEDLINE database and references to the PubMed citations.

Result
An example of the top 100 MedlinePlus terms is listed with text cloud visualization (Fig 1). The link for each term directs users to the previous queries in askMEDLINE database (Fig 2). Preliminary feedback was positive. The responses mentioned that the concept of a bridge Web interface between MedlinePlus and askMEDLINE queries database would benefit both patients and physicians.

Conclusion
A MedlinePlus to askMEDLINE bridge with text cloud visualization adds another resource that may benefit both physicians and patients to find more information on related medical conditions that they might not have considered.

References
Implementation of a Clinical Decision Support Tool to Improve Guideline Compliance in the Prevention of Early-Onset Group B Streptococcus Infections in Neonates

Louise K. Francois Watkins, MD, MPH,1,2 David A. McLean, MD,3 Marion B. Nihen, RN,3 Jonathan M. Wortham, MD,2 Melvin Crum, MS,2 Philipp Carner,2 Mike Jordan,4 Stephanie Schrag, DPhil

1Epidemic Intelligence Service Program, Atlanta, GA; 2Centers for Disease Control and Prevention, Atlanta, GA; 3Lehigh Valley Health Network, Allentown, PA; 4GE Healthcare, Barrington, IL

Early-onset Group B streptococcus (GBS) disease is the most common infectious cause of morbidity and mortality during the first week of life. Approximately 25% of women are asymptomatically colonized with GBS in the genital tract; early-onset GBS disease occurs in newborns when the bacteria are passed from mother to newborn during delivery. Evidence-based guidelines for prevention of early-onset GBS were developed in the 1990s and include recommendations for universal screening of mothers for GBS colonization and administration of intrapartum antibiotic prophylaxis to colonized mothers and those with specific risk factors.1 As a result of widespread guideline implementation, the incidence of early-onset GBS has decreased by more than 80%. Despite this success, errors in guideline implementation have recently been found to occur in more than half of cases of early-onset GBS disease,2 suggesting improved guideline adherence may prevent additional cases of disease. Routine use of a clinical decision support (CDS) tool at the time of patient admission may streamline and optimize guideline implementation.

The Centers for Disease Control and Prevention (CDC) developed an algorithm that uses 13 independent patient variables derived from the guidelines to generate patient-specific, evidence-based recommendations for intrapartum antibiotic prophylaxis for prevention of GBS disease. CDC partnered with Lehigh Valley Health Network (LVHN) in Allentown, PA, an academic medical center performing approximately 4,000 deliveries per year, to create a CDS tool embedded into the Centricity Perinatal electronic medical record (EMR) used at LVHN. Five variables are imported to the CDS tool automatically via existing data fields in the EMR, and a response to one variable, intrapartum screen result, is hardcoded as “unknown” because this testing option is unavailable at LVHN; the remaining 7 variables require entry by the clinician. Values for these variables are then submitted to CDC where they are leveraged by a rule engine (32 rules, modifiable by developers but not end-users) and processed by the algorithm, and a response is generated providing the user with an evidence-based patient management recommendation at the point of care.

The CDS tool was released for use by clinicians at LVHN in August 2014. Users are electronically prompted to employ the CDS tool at the time that electronic admission orders are placed for each patient admitted to the labor and delivery unit. In the first four months after introduction of the tool, there were 1460 total deliveries at LVHN, and the CDS tool was used in 638 cases (44%). Feedback from clinicians was solicited and used to adjust the interface and to tailor response messages to improve their clarity.

The impact of the CDS tool on guideline compliance at LVHN will be assessed after one year of use, with particular attention to areas where compliance may be weaker, such as preterm delivery when GBS screening result is unknown, and antibiotic choice in the setting of penicillin allergy. This project serves as an example of the successful integration of a CDS tool into an existing EMR in order to minimize the time and effort required for clinicians to verify guideline-based recommendations for patient care, and holds important implications for other conditions with complex management guidelines.

References
A Gap Analysis of Competencies and Curriculum in Host-Site Projects of a Field-Based Informatics Fellowship

Laura H. Franzke, PhD, MPH; Sridhar Papagari, MS; Regina Bess, MPA; Herman Tolentino, MD; Brenna Baccaro, MPH; Centers for Disease Control and Prevention (CDC), Atlanta, GA, USA

Abstract
A program overview of an Applied Public Health Informatics Fellowship program that provides a 1-year training at state, tribal, local and territorial health departments. A gap analysis is performed to map and analyze the curriculum and competencies to the host assignment projects. The results will guide how to improve the curriculum and competencies to better serve the informatics needs for health departments.

Background
During 2011, the Centers for Disease Control and Prevention, in partnership with multiple national public health organizations, developed a 1-year Applied Public Health Informatics Fellowship (APHIF) program. APHIF provides informatics graduates the opportunity to participate in an applied training program in public health informatics (PHI) at state, tribal, local, and territorial (STLT) health departments. The fellowship was developed to respond to an informatics workforce shortage among STLT health departments, and to also provide new PHI graduates an opportunity to apply informatics knowledge and skills on host-site informatics projects.

APHIF offers an applied, competency-based curriculum that includes service-learning opportunities; each year APHIF accepts up to ten fellows that are placed at STLT health departments. The PHI competency domains that guide service-learning activities include: (1) analysis, assessment, and evaluation; (2) communication; (3) cultural competence; (4) community dimensions of practice; (5) public health sciences; and (6) leadership and systems thinking. Each fellow is matched to a mentor who provides a learning opportunity to participate in or lead a PHI project. Additional fellow learning opportunities include: orientation, monthly seminars, communities of practice (COP), and peer-to-peer learning on various public health informatics topics. The host projects are different in scope and skill requirements, allowing fellows to be matched to their area of expertise and learning needs. Examples of host site assignments topics are as follows: electronic lab reporting, meaningful use, interoperability and/or electronic health records. Since 2012, (n=27) fellows have been placed in 15 states.

Methods
A two-fold gap analysis will be performed for the APHIF program to compare between actual and desired performance. First, the APHIF curriculum will be mapped to the PHI competencies, and second, the host site projects will be mapped to PHI competencies. The mapping and analysis will examine current to desired fellowship performance state; this will elicit potential gaps and strengths in the curriculum, competencies, and host projects. This gap analysis will guide how the curriculum and competencies can be improved to more fully support the STLT health department’s informatics needs. Additionally, this gap analysis will guide mentors and supervisors about how to potentially improve host assignment projects to provide additional learning opportunities.
Abstract:

The current means for reporting the door to healthcare provider time of patients is neither accurate nor reliable. We have created a reliable method of automatically capturing this time using Bluetooth-based technology that does not hinder the Emergency Department (ED) workflow. The results show accurate measures, within seconds, of patient wait times and reliable detection of providers within key locations, without negatively impacting the ED workflow.

Introduction:

The Center for Medicare and Medicaid Services (CMS) requires that EDs log and report certain performance measurements including the door to healthcare provider time for the patients. Current data collection methods for time performance measures that EDs employ are not automatic in nature and require extra work for the provider to document the time a patient is seen. Our work seeks to provide an automatic method of provider time documentation, while not interfering negatively with the ED workflow.

Methodology:

Using the wireless sensor technology, single board computers with Bluetooth detection and automatic time correction software were installed on the ceilings above the hospital beds in the center of ED rooms. These computers detect when a Bluetooth beacon (a broadcaster) is within the measuring range and calculate the distance between itself and the beacon. This distance is used along with the accepted room radius to determine when a healthcare provider (carrying the beacon) is in an ED room seeing a patient. If the healthcare provider is indeed in the room, then the clinician time for the door to clinician time is logged. The door to clinician time is calculated from subtracting the door time, which was collected at the time of patient data registration, from the clinician time. Clinicians’ measure of handwritten times were compared against our detected beacon times.

Results and Discussion:

A total of 17 out of the 500+ clinician–patient encounters (to this date) have been reviewed, of which six set of data were excluded because the care was provided by resident trainees, who were not given the beacons. The overall results show that the recorded door to clinician times were accurate where the beacons were detected and the times recorded in seconds. The times recorded by the software were compared with time.gov for accuracy and also compared against the clinicians’ handwritten times, but the logged times are more accurate.

Conclusions and Future Research:

As our preliminary results show in the ED, this inexpensive Bluetooth broadcasting technology offers an automatic and reliable way to track provider-patient encounters accurately to the second. We hope to extend the study to include assessment of team interactions, given the installation of additional sensors in areas such as over the hallway patient beds and the Triage area.

References:

An Examination of Standalone Personal Health Record Use by Patients with Type 2 Diabetes

Kevin T. Fuji, PharmD, MA1,2, Amy A. Abbott, PhD, RN1,3, Kimberly A. Galt, PharmD, PhD1,2

1Creighton University Center for Health Services Research and Patient Safety, Omaha, NE; 2Creighton University School of Pharmacy and Health Professions, Omaha, NE; 3Creighton University College of Nursing, Omaha, NE

Introduction
The use of personal health records (PHRs) by patients has been hypothesized to improve patient knowledge about their medical conditions, aid in maintaining comprehensive health records, and facilitate patient engagement in their own care.1 PHRs hold particular promise for patients with chronic conditions, such as diabetes, who often manage multiple medications and co-morbidities.1 However, there is limited information on the actual impact of this technology on clinical outcomes and self-care behaviors. The purpose of this study was to examine the impact of standalone PHR use on hemoglobin A1c (HbA1c) and social cognitive outcomes of patients with type 2 diabetes.

Methods
This pilot study utilized a randomized controlled trial in which 140 participants were assigned either to a PHR group who used Microsoft HealthVault to manage their diabetes-related health information, or to a usual care control group. Outcomes were assessed at baseline and 3-6 months later. HbA1c levels were obtained from medical records, while social cognitive outcomes were measured via a patient survey: 1) modifying factors to disease self-management (perceived barriers to self-management; diabetes education; and social support); 2) diabetes knowledge; and 3) self-efficacy. A score of 1-5 was calculated for each concept, with higher scores indicating a more positive response. Interviews were also conducted at 3-6 months with all PHR group participants, and analyzed using an iterative thematic analysis.

Results
A total of 117 participants completed the study (61 control, 56 PHR). No significant change in HbA1c was observed in either group (Control: 7.53%±1.53% to 7.75%±1.22%, p=.252; PHR: 7.86%±1.96% to 7.98%±2.01%, p=.535). No significant changes were observed in the control group for social cognitive outcomes: modifying factors (3.78 to 3.70, p=.204); self-efficacy (3.77 to 3.80, p=.317); and diabetes knowledge (4.13 to 4.19, p=.254). In the PHR group, no significant changes were observed in modifying factors (3.78 to 3.83, p=.297) or self-efficacy (3.82 to 3.83, p=.657); however, there was a significant improvement in diabetes knowledge (4.11 to 4.25, p=.029). Five themes emerged from the thematic analysis: 1) few people use the PHR more than once or twice; 2) PHR users find value in self-management of their health; 3) PHR users are persons who are already tracking information; 4) PHR users do not proactively share the information with their providers; and 5) providers who knew patients were trained in PHR use did not ask for the information.

Conclusion
Short-term use of a standalone PHR improved patients’ diabetes knowledge, but failed to have a significant effect on HbA1c and other social cognitive outcomes. While the PHR has value for self-management, PHR users were already using existing tools and did not engage in sustained PHR use. In order for PHRs to help patients move from enhanced diabetes knowledge to improved clinical and self-management outcomes, it must demonstrate clear value beyond existing health information tracking tools and promote information sharing between patients and providers. Future PHR design could incorporate patient-centered decision support tools (e.g. preventive care reminders or identification of potential drug interactions) that provide value beyond use as an information repository. Providers should also consider introducing the PHR as a self-management tool for patients during diabetes education.

References
Therapeutic Area and Research Use Case Based Data Profiling & Quality Assessment Framework

Purav Gandhi, MBBS, MBA\(^1\), Jasmine Avasia, BSc\(^1\), Shrestha Anantha, BE\(^1\), Rajinder Sobti, BTech, MBA\(^1\)

\(^1\)ConvergeHEALTH by Deloitte, Newton, MA

Abstract

We have developed a framework for data profiling and quality assessment for real-world evidence that performs an overall data quality assessment, therapeutic area, and research objective based evaluation. The profiler leverages definitions and rules from a content library which can be updated as understanding of data quality evolves, and the findings presented in a highly intuitive user interface to simplify data due diligence and quality assessment process.

Introduction

Data quality is important in real-world evidence based research using clinical and claims data, to determine the feasibility and level of confidence in the results of a study. As our understanding of secondary uses to data improves, we recognize that data quality is not only measured in quantitative terms such as number of patients and longitudinality of the data set, but now has a lot of qualitative considerations such as comprehensiveness across diagnosis, treatment, procedures, and similar areas. Also, clinical nuances vary across disease areas, such as specific biomarkers to be tracked for research. Currently limited number of such intelligent data profiling solutions exist that can profile and evaluate data quality based on therapeutic area of interest.

Method

We have designed a unique data profiling and quality assessment framework (Data Asset Explorer) that helps assessment of data quality at three different tiers: (1) General data profiling and quality assessment, which includes validation of data against metadata like number of rows, longitudinality, overlapping of fields across diagnosis, treatment, procedures, etc., (2) Therapeutic area based profiling and assessment, that includes validations such as data availability for tests and biomarkers related to the therapeutic area, number of follow-ups for specific tests, and disease specific longitudinality threshold, and (3) Research objective specific profiling, which evaluates the presence of specific data fields relevant for the research objective (e.g., presence of cost data for economic impact analysis), and the data density in those fields.

Solution

Data Asset Explorer profiles the data across different data sources through a generic and extensible metadata framework, Common Meta Data Model (CMDM). It is extensible to support different types of healthcare data source as indicated in the diagram. Query Generator leverages the data quality rules and profiles from Content Library (central repository of the rules, measures and models) and CMDM to query the specific data sources and also compare different data sets. The results of the profiling and quality checks are rendered through a visual interface.

Discussion

Data Asset Explorer will help the researchers navigate through the snapshot (both generic and targeted) of different sources for evaluation and comparison purposes before. This will generate efficiency in data due diligence and feasibility assessment process for various studies. We are in the process of testing the framework for extensibility to other data sources such Omics data, as well as its ability to capture comprehensive rule set for every disease area.
Dashboard Visualizations of Emergency Department Throughput Metrics

Swaroop Gantela, MD, Todd R. Johnson, PhD, Nnaemeka G. Okafor, MD, MS, David J. Robinson, MD, Amit M. Mehta, MD, Charles L. Maddow, MD, Brent R. King, MD, Tina E. Chacko, PA, Salsawit Shifarraw, BBA, Vickie Nguyen, MS, Adriana Stanley, BS, Amy Franklin, PhD

1 University of Texas Health Science Center Houston, Houston TX; 2 Nemours/Alfred I. duPont Hospital for Children, Wilmington DE; 3 Memorial Hermann Healthcare System, Houston TX

Introduction

The Centers for Medicare and Medicaid Services are exploring the use of payment determination measures that will include Emergency Department (ED) throughput metrics. As timely and effective care becomes defined by (non)adherence to goals for care processes, such as overall length of stay and time from admission decision to departure, providers and administrators need a way to improve efficiency. We discuss the use of visualizations of throughput data to provide cognitive support for real-time comprehension and projection of the state of ED care processes. Borrowing from Pimentel and Barrueto’s application of statistical process control (SPC) methods in understanding historical ED processes, we propose that SPC visualizations of real-time data can provide visual statistical cues that may more accurately capture true process performance and potentially provide early warnings of process change.

Methods

SPC visualizations support recognition of the normal variation inherent to a process and “special cause” variation resulting from possible statistically significant changes in process behavior. For example, in Figure 1, the SPC chart indicates a care process exhibiting normal variation until the latter points where a specific trend criterion is met (10 points out of 12 above the median), suggesting a special cause or process change. Even though many points prior to this are above the target, those points are not statistically different from the median. In contrast, the goal-based chart in Figure 2 highlights all over-target values, which encourages treating all variation as special cause—an approach shown to increase process variation.

Example Metric: Time from Provider Assigned to Disposition Selected

<table>
<thead>
<tr>
<th>Goal: less than 120 minutes</th>
</tr>
</thead>
</table>

**Figure 1.** Statistical Process Control chart  
**Figure 2.** Goal-based visualizations

Conclusion

SPC methods have been long used in mechanical production environments to assess historical process performance and detect real-time deviations from control. While the ED environment is less deterministic, highlighted visualizations of control charts can play a supportive role in ED situation awareness and decision-making.

References

Electronically Collecting Nocturnal Heart Failure Information

Cubby L. Gardner, MS, RN, FNP-C1, Harry B. Burke, MD, PhD2
1Daniel K. Inouye Graduate School of Nursing, 2School of Medicine
Uniformed Services University of the Health Sciences, Bethesda, MD

Background

The current approach to the outpatient management of heart failure combines patients’ retrospective self-reports of symptoms and activity with a physical examination and laboratory results during a routine office visit. Clinicians expect that patients will accurately recall significant changes in symptoms and activity since the last clinic visit. Unfortunately, there is a great deal of variability in patient’s self-reports and in subjective assessment instruments, recall is, many times, not veridical with past events and it may be especially difficult for heart failure patients because of the prevalence of mild cognitive impairment due to both heart failure itself and advancing age. Thus, retrospective objective and subjective information recalled at clinic visits may not provide the accurate, complete, and detailed information that clinicians require for the optimal management of their patients. We provided patients with physiologic devices to measure their clinical status at home. We hypothesized that patients would be able to use these devices overnight in their homes and that this would improve the information available to clinicians.

Methods

Patients were trained on five objective and subjective data acquisition devices (weight scale, blood pressure monitor, pulse oximeter, actigraph, and an iPad-based subjective question set). They used them to collect data at home, overnight for up to six nights. The raw data was downloaded using device-specific software, cleaned using R statistical language, and uploaded into the MySQL study database. We expected patients to use the scale once a night, BP and iPad twice a night, and the continuous devices all night for 8 hours. The ratio of observed over expected (O/E) device use, where 1.0 is observed = expected, was calculated and tested using the Student’s t-test.

Results

Participants were 39 clinical heart failure patients, mean age 68.1 (SD, 12.3), 71.8% male, New York Heart Association functional class I, 10 patients; class II, 22 patients; and class III, 7 patients. Three of the devices were used as expected (Table). The scale was used significantly less than expected O/E 0.80, p<0.001. The pulse oximeter, worn on the wrist overnight to collect heart rate and blood oxygen saturation, was worn as expected O/E 0.90, and was marginally significant, p=0.05. The actigraph, which was worn on the wrist and measured activity, was used significantly more than expected, O/E 1.93, p<0.001.

Table. Patient device use.

<table>
<thead>
<tr>
<th>Episodic devices</th>
<th>Days used Mean (95% CI)</th>
<th>Range</th>
<th>Average number measurements taken</th>
<th>Observed</th>
<th>Expected</th>
<th>O/E^1</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale^2</td>
<td>4.50 (4.02, 4.98)</td>
<td>2 - 6</td>
<td>4.87</td>
<td>6</td>
<td>0.80</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Blood pressure monitor</td>
<td>5.58 (5.22, 5.95)</td>
<td>1 - 6</td>
<td>12.54</td>
<td>12</td>
<td>1.03</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>iPad questions</td>
<td>5.56 (5.14, 5.99)</td>
<td>0 - 6</td>
<td>13.58</td>
<td>12</td>
<td>1.10</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Continuous devices</th>
<th>Days used Mean (95% CI)</th>
<th>Range</th>
<th>Time used (hours)</th>
<th>Observed</th>
<th>Expected</th>
<th>O/E^1</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulse oximeter</td>
<td>5.51 (5.20, 5.82)</td>
<td>3 - 6</td>
<td>43.71</td>
<td>48</td>
<td>0.90</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Actigraph^3</td>
<td>5.79 (5.56, 6.00)</td>
<td>2 - 6</td>
<td>95.58</td>
<td>48</td>
<td>1.93</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

^1Ratio of observed over expected (O/E) device use where 1.0 is observed equals expected.
^2The results are based on the 24 patients who used the scale.
^3Three patients used a Zephyr device for activity; they are not included in the activity results.

Discussion

We found that the labor-intensive scale, which had to be set-up for each measurement, was used less than expected, whereas the actigraph, which is simply placed on the wrist, was used more than expected. In another study we assessed patient perceptions of device usability and found high acceptance. This study is important because we need to know which devices patients will actually use in their home and why they will use them.

The views expressed are those of the authors and do not reflect the official policy or position of the Uniformed Services University of the Health Sciences, the Department of the Defense, or the United States government. All devices met Protected Health Information (PHI) and Personal Identifiably Information (PII) requirements. WRNMMC IRB Protocol number 400924-4.
Evaluating the Effect of a Nursing Flowsheet Merge on Clinical Nursing Documentation Efficiency in a Research Hospital

Kyungsook Gartrell, RN, PhD1, 2, Helen S. Mayberry, MSN, RN2, Michael Krumlauf, RN, BSN2, Michael Nansel, RN, BSN3, Minnie Raju, MS, RN3, Gwenyth R. Wallen, RN, PhD2, Caitlin W. Brennan, PhD, APRN2

1National Institutes of Health/National Library of Medicine, Bethesda, Maryland; 2National Institutes of Health/Clinical Center Nursing Department, Research & Practice Development Section, Bethesda, Maryland; 3National Institutes of Health Clinical Center, Department of Clinical Research Informatics, Bethesda, Maryland

Introduction

The National Institutes of Health (NIH) Clinical Center has used the clinical research information system (CRIS) to support the documentation of clinical care and biomedical research since 2004. CRIS is an amalgamation of vendor-based applications that encompass the NIH electronic health record (EHR). The core system in CRIS is the Sunrise 14.3 application suite. CRIS is used by physicians, nurses, and ancillary and administrative staff in performing a variety of information tasks related to patient care. Tasks include 1) managing registration, admissions, and discharges and transfers, 2) maintaining demographic data and clinical research protocol information, 3) providing bed management, 4) attributing medical activities (orders, appointments and documents) to specific research protocols, 5) retrieving laboratory and radiology results, 6) documenting care plans, vital signs, medication administration, intake and output, and progress notes, and 7) reviewing outside documentation. Approximately 90% of the NIH Clinical Center’s medical record content is generated from CRIS. The nursing outcomes management team and the department of clinical research informatics routinely collaborate, with a goal of using CRIS data to inform quality improvement and patient safety efforts.

One problem with nursing documentation in the CRIS was that nursing assessment and intervention data were documented in separate screens, making information difficult to find and to document. Little is known about best practices for use of nursing data in EHRs to communicate plans of care among multiple professions. The significance of this evaluation is that it serves as the necessary first step in describing the contributions nurses at NIH make in translational research and how other members of the interprofessional team use nursing information in patient care. In December of 2014, the content from two nursing flowsheets was merged into one flowsheet to better reflect the continuity of care provided by nurses. The aim of this investigation is to evaluate the effect of the nursing flowsheet merge on clinical nursing documentation efficiency.

Methodology

A retrospective query from CRIS will be used to extract data for six months pre- (June-November 2014) and six months post- (January-June 2015) implementation of the nursing flowsheet merge. The effect of the nursing flowsheet merge will be measured by change of frequency of usage between pre- and post-data in several categories. Nursing assessment and intervention information is documented in 21 sections in CRIS that can be classified into 5 categories (physical, psychosocial, patient education, activity, and safety). Nursing documentation will be quantified as the total number of saved and modified notes in each category. Frequency of usage will be calculated by total number of times specific categories of CRIS were accessed to view and read information.

Evaluation Results and Conclusions

SPSS will be used for statistical analyses. Descriptive statistics will include frequencies and percentages to describe the total number of nursing notes saved and modified and the total number of times nursing assessment and intervention notes were accessed for viewing purposes. Chi-square tests will be used to investigate whether distributions of nursing assessment and interventions differ between pre- and post-implementation. Each of the aforementioned 5 categories of nursing documentation will be presented for the pre-and post-implementation comparison. If a statistically significant difference is detected between inpatient and outpatient units, the data will be reported separately. Expected results are that increased efficiency of nursing documentation occurred as a result of the flow sheet merge and is quantified as an increase in the number of saved and modified nursing notes, and by an increased number of viewed nursing notes.
Development and Validation of a Measure for EHR Related Unintended Consequences with Direct Care Registered Nurses

Sheila M. Gephart, PhD, RN; 1 Judy Dye, MS, NP; 1 Alycia Bristol, MS, CNS, RN; 1 Brooke A. Finley; 1 Jane Carrington, PhD, RN 1
1 The University of Arizona, Tucson, AZ

Introduction
Electronic Health Records (EHRs) are being widely implemented with promise to support information availability, coordination of care and clinical decision making. Yet, the swift implementation of these imperfect systems may threaten patient safety as clinicians, including nurses, devise EHR workarounds. Guided by previous qualitative studies, a quantitative measure (i.e. CG-UCE-Q) was developed to measure the impacts and prevalence of unintended EHR consequences. The purpose of this study was to determine the extent to which the measure is valid and reliable then use it to describe nurses’ frequency of experience with EHR related unintended consequences.

Methods
Nurse Informaticist (NI) experts were recruited and electronically completed a survey identifying each item’s relevance on the CG-UCE-Q. Data was analyzed using descriptive statistics and items were grouped based on: 1) relevant or relevant with minor revision; and 2) unable to assess relevance without revision or irrelevant. Reviewers suggested improvements and 2 items were added for round 2. Direct care R.N.s were then recruited from a single health system with 2 separate hospitals to complete the CG-UCE-Q. The same survey was distributed 2 weeks later and responses to the two testing times were linked. Psychometric testing of reliability and validity was then conducted. Qualitative data was also collected and is reported elsewhere. The measure is scored by assessing the frequency with which an UC occurs on a scale from 0-6 (0= never, 1= once a year, 2=once every couple of months, 3= once a month, 4= once a week, 5= once a shift and 6= multiple times a shift).

Results
NI experts (N=5) ranged from moderately to extensively expert. After two rounds of testing, the content validity index was 0.96 and 44 items were included (2 added based on experts’ recommendations). Then, of 1100 RNs invited to participate, 226 responded and were screened for eligibility, 113 eligible RNs completed the first survey. While 87 completed the second survey, after removing incomplete responses, 68 were usable for test-retest measurement. R.N.’s worked in various acute care settings, although predominantly represented medical-surgical (26%) and intensive care (32%). Ages varied widely, 91% were female, and 74% were B.S.N. prepared or greater. The CG-UCE-Q was highly internally consistent (Cronbach’s α = 0.94; ICC=0.91) with strong item-total correlations (0.30-.80) for the majority of items. Individual responses changed over time ($X^2 = 980 (39), p < 0.001$) and did not support test-retest reliability, although median scores across the groups were highly similar over time. Highest item-total correlations were identified for threats to patient safety when attempting to assess risk for complications (0.79), when coordinating care at discharge (0.79), transferring patients within the hospital (0.75), and when coordinating care within the hospital (0.75). Highest frequency of UCs (i.e. at least once a shift to multiple times a shift) were identified for increased workload (M=4.9, SD=1.5), interruptions requiring task shifts away from the computer (M=5.55, SD=0.77), changing workflow (M=4.4, SD=1.8), and duplicative documentation (M=4.9, SD=1.8).

Discussion
This is the first, to our knowledge, quantitative measure of nurse experience with EHR related UCs. Nurses experience UCs frequently and indicate they threaten patient safety in care at least once a shift to once a week. The highest risk relates to poor information availability or EHR misrepresentations that occur during transitions and when care coordination and medication administration is an issue.

Conclusion
Psychometric testing supports the internal consistency and content validity of the CG-UCE-Q. Test-retest reliability demonstrated that the measure is not stable, but may reflect a changing nurse experience over time. Construct validity testing is underway. Of most concern, nurses reported high frequency of UCs related to patient safety, particularly at times when coordination and transitions in care occur.

1482
Developing Principles and Best Practices for Structured Documentation

Emily J. Gesner, DNP, RN-BC\textsuperscript{1}, Sarah A. Collins, PhD, RN\textsuperscript{1,2,3}, Perry Mar, PhD\textsuperscript{1,2,3}, Roberto A. Rocha, MD, PhD\textsuperscript{1,2,3}  
\textsuperscript{1}Clinical Informatics, Partners eCare, Partners Healthcare System, Boston, MA;  
\textsuperscript{2}Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA;  
\textsuperscript{3}Harvard Medical School, Boston, MA

Introduction

Electronic Health Records (EHRs) that are highly customizable to increase end-user satisfaction are considered desirable by most health care systems [1]. Customization of EHRs offers an opportunity to optimize the system to meet a variety of clinical, administrative, and policy needs across an institution or enterprise. However, configuring the system without a set of principles and best practices to guide accurate and consistent data capture is a significant challenge for any organization.

Standardizing clinical data definitions in an EHR increases overall data consistency [2]. Defining how clinical data is structured within an EHR is constrained to the actual functionality of the system. The aim of this project is to (a) describe the principles and best practices for structured data capture in documentation forms, and (b) to conduct a preliminary evaluation of a sample of documentation forms to assess how well they meet the defined principles and best practices.

Methods

A workgroup was formed and began with an iterative process, including basing decisions on data standards such as HL7 and other prominent informatics principles.

After defining the preliminary principles and best practices, it was determined that a baseline assessment needed to be completed for a sample of existing forms. From a recent set of submissions to the workgroup for locally managed structured data elements, a sample of 44 structured documentation forms consisting of a total of 310 structured data elements was chosen to complete this task.

During the review, each form was analyzed using the pre-defined principles and best practices. Principles and best practices that were considered “Not Applicable” to the documentation forms being assessed were omitted from the analysis.

Results

The effort of the workgroup resulted in a preliminary set of 7 principles for building structured data elements. It was determined that one or more best practices could be related to each principle. For example, the principle “Promote reuse and sharing of data elements” can be related to the best practice of “If a structured data element has the same function on multiple forms (e.g. documentation of pain level), then it is best to try to reuse the same structured data element on all the structured documentation forms.”

When the baseline assessment was completed, the “avoiding repeating details” and “using singular naming” instead of plural was met 100% of the time. The principles of “using a structured data element”, “use an additional smart data element with specific details”, and “removing “other” from category lists” were met 0% of the time.

Discussion

The definitions of principles and best practices for structured clinical documentation should be clear and concise. After completing the baseline assessment, the effort to obtain consistent data capture within the structure data documentation forms is a complex problem in which solutions will need to evolve over time. The results are extreme due the preference to build the structured data forms by user preference over quality data is an issue that cannot be immediately fixed. Also building to the specifications of the end-user for a convenient UI is considered a significant aspect of the problem. In the short term, this line of work will continue to evolve including recommending specific data requirements for consistent data capture and sharing data elements to improve data consistency and quality over time.

References


2. Los RK. Supporting Uniform Representation of Data: Structuring Medical Narratives for Care and Research. Erasmus University; 2006.
Streamlining Access to Cancer Data - An Institutional Experience

Leanne Goldstein, DrPH, Rebecca Ottesen, MS, Kelli Olsen, MS, CTR, Janet Nikowitz, CTR, Joyce Niland PhD; City of Hope, Duarte, CA

In this Big Data-era, the lack of large standardized databases of clinical data on all cancer patients represents a void that can be considered “criminal”. City of Hope (COH) has embarked on the creation of such a Disease Registry of standardized clinical data for any cancer site, to support genotype-phenotype analyses, tissue correlations, outcomes, and quality of care research. Our approach to protocols, privacy, security information modeling, and data integration will be described.

When even the tiniest companies can track every piece of data, predict their trends, analyze cohorts and monitor their progress, why can’t we create such a database for cancer patients, to allow us to learn from patients’ experiences and make more educated treatment decisions to extend and improve lives? In a time where access to data is crucial for making decisions about clinical care, COH is streamlining the process of obtaining data used for research and informing treatment decisions through the creation of an institutional Disease Registry. Such a registry is critical to make new discoveries towards precision medicine, conduct genotype-phenotype research, tissue correlative analyses, outcomes research, and quality of care assessments. In many institutions, it can be difficult to obtain research study approval by the Institutional Review Board (IRB) because the submitted protocol may come back with concerns regarding data security or other issues that often delay research. An additional layer of complexity is that data for outcome and tissue correlative studies are typically difficult to obtain because the information is often collected in various data silos across the institution. Principle Investigators (PIs) tend to utilize their fellows and residents to collect the data elements required for their research. This information is typically entered on to unreliable and unsecured spreadsheets, and at the end of the study the valuable information abstracted is often never used again. All of this is disadvantageous to patients who are in dire need of obtaining modern life-saving treatments based on current research and data.

The COH Disease Registry program falls under an IRB-approved “umbrella” protocol which dictates the infrastructure for collection of data for all cancer disease sites, describing the data collection scope, security, privacy, honest broker, and data integration processes. Each disease site will be supported by a linked “spoke” protocol, as defined by disease-specific research teams, to allow the collection and access to pooled institutional data, including information collected as codified elements in the Electronic Medical Record (EMR) as well as data abstracted by the Cancer Registry. We aim to use common data model CDM version 5 in OMOP format. Existing data such as data from pathology synoptic reports are integrated into the Disease Registry to allow physicians to have codified data from the time of diagnosis. Natural language processing is being used to facilitate the abstraction of data elements by Disease Registry Systems Analysts of information found in unstructured text. When additional data for a research study are required, it can be abstracted from dictations in the EMR such as clinical notes or MRI scans. Detailed data beyond the elements in the EMR and Cancer Registry are entered into a secure REDCap database. All Disease Registry data across cancer sites will be integrated via the Enterprise Data Warehouse. It will also be incorporated into a data mart for our internal i2b2 instance so that users across the institution will have enhanced capabilities for refining their queries and targeting the specific cohort or specimens they wish to use in research. A library of oncology based queries will be developed that could be used across institutions.

The Disease Registry program will focus on collecting a variety of oncology based data elements and optimizing tools for data abstraction. We aim to coordinate the effort by PIs interested in specific diseases, approval through the IRB, enhancements to the security of data collection, and retention of study specific data for future uses thereby reducing duplicative and redundant efforts and wastes of research money and resources. Ultimately the Disease Registry program will allow us to maintain an open and collective institutional data repository that will generate knowledge and research in order to improve the lives of cancer patients treated at the COH. Furthermore, the Disease Registry will serve as the vehicle to capture and export the clinical / phenotypic data required for our new membership in the national Oncology Research Information Exchange Network (ORIEN).

1 LA Times Editorial, Laurie Becklund, February 2015
A Tale of Two Layouts: Vignette versus Structured Interview for Layout of an Electronic Handoff Tool

Nathan C. Gonzalez,1 Ray McKelvy,1 Karol L. Fleming MBA, JD,1 William B. Webb BSPH, MBA,1 Robert L Shapiro MD, MSCI1,2

1 MEDarchon, Inc. 113 29th Avenue South, Nashville, TN 37212
2 Cardiovascular Division, Washington University in St. Louis School of Medicine, 660 S. Euclid Avenue, St. Louis, MO 63110

Background: The benefits of an electronic handoff tool on patient care are known to be significant. However, our review of the clinical literature has found no studies focusing on the usability of such a tool. Poor usability may limit adoption and gains of an e-handoff tool and the present study sought to define the proper layout (user experience) to support current practice.

Methods: Doctors and nurses with at least two years of clinical experience were selectively recruited based upon their role, specialty and patient population. After giving informed consent and completing a pre-study survey about their practice and handoff tendencies, subjects were asked to narrate a clinical handoff without disclosing PHI. They then were presented with a number of visualizations for handoff data and asked to arrange the factors on an imaginary computer screen. These responses were codified and summarized using descriptive statistics and the Z score.

Results: 24 subjects participated in the study; however, 1 participant was paged and had to drop out. 23 subjects completed the study and produced data that could be analyzed. Figure 1 shows the relative position of the various factors in the narrated handoff with 95% confidence intervals. The majority of participants created a screen layout with either two (43%) or three (39%) columns. In subanalysis, those internal medicine and medical subspecialty physicians were more likely to prefer a three column format (0.7 vs. 0.2; p=0.012). No other subgroups showed a significant difference in preference although nurses trended toward preferring a less complex column displays (0.86 vs. 0.50; p=0.107). Figure 2 shows the relative position of the various factors when arranged by the participants. While no measures reached significance in subanalysis, precautions tended to be placed closer to the end by surgeons (p=0.13) and those clinicians caring for pediatric populations (p=0.16); patient status was placed closer to the end by physicians (p=0.26); and a summary of clinical events was placed closer to the end for ICU-based clinicians (p=0.21). Conclusion: Our study demonstrates that the user experience for a handoff tool should be customized based on the user’s role (MD vs. RN), specialty/area of practice and patient population (adult vs. pediatric). It also suggests the visual flow of information for these various roles although this remains to be proven in an upcoming validation study.

Figure 1: Oral Vignette Factor Order
Figure 2: Layout Exercise Factor Order

Figure 1 shows the relative chronological appearance of factors in an unprompted oral handoff exercise while Figure 2 shows how clinicians ordered the factors in a prompted geospatial exercise. The color coding of the dots is as follows: Light Blue: Demographics; Dark Green: Medical History; Dark Blue: Physical Findings; Orange: Medications; Light Green: Precautions; Red: Patient Status; Purple: Interpretation and Care Planning.

Integrated Clinical Decision Support Systems: Systematic Review and Classification of Online Medical Calculators

Tim A Green, BS¹, Chi-Ren Shyu, PhD¹,²,³,⁴
¹Informatics Institute, ²Computer Science Department, ³School of Medicine, ⁴Electrical and Computer Engineering Department, University of Missouri, Columbia, MO

Electronic capture of clinical data enables a variety of clinical decision support systems (CDSS). The recent proliferation of EMR’s raises the expectation that new predictive models and other types of decision support will be developed at a more rapid pace than ever before. Making clinical decision support systems available at the right time within the appropriate clinical workflow will be even more important as the amount of CDSS increase. Medical calculators are an important class of CDSS. These calculators embody evidence based medicine and are generally based on scientific literature[1]. They come in a variety of formats[2], and can even be patient facing in nature[3]. A systematic study of the integration of clinical decision support systems into EMR workflows will provide valuable insight into the best methods to deliver new evidence based medical recommendations. The proposed plan of study consists of several phases leading to a better understanding of medical calculator workflow.

<table>
<thead>
<tr>
<th>Source 1 (n=138)</th>
<th>Source 2 (n=498)</th>
<th>Source 3 (n=126)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Result Type</strong></td>
<td><strong>Count</strong></td>
<td><strong>Percent</strong></td>
</tr>
<tr>
<td>diagnosis</td>
<td>2</td>
<td>1.45</td>
</tr>
<tr>
<td>recommendation</td>
<td>13</td>
<td>9.42</td>
</tr>
<tr>
<td>advice</td>
<td>29</td>
<td>21.01</td>
</tr>
<tr>
<td>probability</td>
<td>23</td>
<td>16.67</td>
</tr>
<tr>
<td>classification</td>
<td>75</td>
<td>54.35</td>
</tr>
<tr>
<td>discrete data</td>
<td>33</td>
<td>23.91</td>
</tr>
</tbody>
</table>

Table 1 – Methods of presenting calculator results from most to least actionable probability, classification, and discrete data representing progressively less actionable results. As Table 1 illustrates, more than 75% of the calculators from each of the three sources return results in one of the bottom three result types. A prior study by Kawamoto et al. [4] indicates successful adoption is positively correlated with provision of recommendations (top three results types) rather than assessment (bottom three result types). Our findings indicate that the majority of available calculators do not fall into the categories associated with successful adoption. Perhaps development of future medical calculators should account for the presentation of results in ways that promote action and adoption, rather than simply presenting a score or calculating discrete data.

<table>
<thead>
<tr>
<th>Yrs experience</th>
<th>Wald (z)</th>
<th>Probability of use</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-6</td>
<td>4.03</td>
<td>0.919</td>
<td>0.777</td>
</tr>
<tr>
<td>7-10</td>
<td>-0.94</td>
<td>0.349</td>
<td>0.833</td>
</tr>
<tr>
<td>11-20</td>
<td>-1.84</td>
<td>0.066</td>
<td>0.722</td>
</tr>
<tr>
<td>&gt;20</td>
<td>-3.6</td>
<td>0.00032</td>
<td>0.486</td>
</tr>
</tbody>
</table>

Table 2 – Correlation of experience to probability of using medical calculators

In the current phase, we surveyed clinical users and analyzed three web-based medical calculator services. Meta data were collected from these services for further analysis. Of particular interest was the discovery that the bulk of medical calculators assessed delivered results to the user in passive forms, versus active, actionable formats (Table 1). A result type of diagnosis represents the most actionable format, with recommendation, advice, and probability, classification, and discrete data representing progressively less actionable results. As Table 1 illustrates, more than 75% of the calculators from each of the three sources return results in one of the bottom three result types. A prior study by Kawamoto et al. [4] indicates successful adoption is positively correlated with provision of recommendations (top three results types) rather than assessment (bottom three result types). Our findings indicate that the majority of available calculators do not fall into the categories associated with successful adoption. Perhaps development of future medical calculators should account for the presentation of results in ways that promote action and adoption, rather than simply presenting a score or calculating discrete data.

Surveys were sent to all practicing physicians at an academic medical hospital, and data from the 108 responses have been tabulated. Findings from the survey indicate that years of experience is predictive of the probability of use of medical calculators (Table 2), with the Wald statistic for the 1-6 and >20 categories significant at α=0.05. The probability of use for years 1-6 is 0.919, whereas the probability for years > 20 is just .486. This suggests that awareness may be low among more experienced clinicians, and/or they may not see a need for these calculators. Survey free text responses suggest that experienced practitioners rely on experience, have trainees use them, or were unaware of them.

References

Variation Among Providers in Cost of a Knee Replacement Episode

Thomas R. Gurgiolo, MA¹, Taylor R. Pressler, MS¹, Scott Orme, PhD¹, Robert Kershner, MPT MBA², Jonathan D. Mathieu, PhD²,

¹Aver Informatics, Columbus, OH; ²Center for Improving Value in Health Care, Denver, CO

Abstract
Payment reform will include bundled payments. CIVHC is an all payer claims database within Colorado and presents a unique opportunity to compare the costs and utilization patterns within an episode of care between providers. The proposed poster will compare results of a standard Knee Replacement episode definition for Medicaid patients by provider facility. This poster presentation will examine geographic variations in cost and utilization within this episode of care.

Introduction
Bundled payments are defined as the reimbursement of health care providers “on the basis of expected costs for clinically-defined episodes of care”¹. In 2013, Medicare started a pilot bundled payment program. The Center for Improving Value in Health Care (CIVHC) is an all payer claims database for Colorado and offers a unique opportunity to examine claim and quality variances in care between different providers within the state.

Methods
The Knee Replacement episode was constructed using the Prometheus 3.6 definition. This definition provides codes and definitions for establishing which claims are part of a valid Knee Replacement episode, including which claims are considered Typical or Potentially Avoidable Complications (PAC). Episode results were compiled using the Aver Informatics platform and the results stratified by blinded facility provider. Results will be evaluated using a Student’s t-test and an adjustment for multiple comparisons.

Results
In the CIVHC data set there were a total of 2,108 valid Knee Replacement episodes, at an average cost of $10,270. Of this cost, an average of $8,173 were identified as typical costs and $2,079 were identified as potentially avoidable costs. Among providers with at least 30 episodes, the total average cost of a knee replacement varied from $6,469 to $16,222; a sample of the payer details is shown in Figure 1 below. Additional results, including analysis by region, will compare the results of this episode of care at the time of presentation.

Figure 1.

Conclusion
Using bundled payment methodology, a wide variety of total cost have been identified for Knee Replacement episodes in the CIVHC Medicaid population. Further analysis of the reasons for this variation may result in opportunities for optimizing patient care and payment reform.

References
Identifying Children with Technology Dependence through Use of Administrative Data

Baria Hafeez¹, Babitha Haridas², Anup D Patel²,³, Zachary Grinspan¹,⁴,⁵
¹Department of Healthcare Policy and Research, Weill Cornell Medical College, New York, NY
²Division of Pediatric Neurology, Nationwide Children’s Hospital, Columbus, Ohio
³Department of Clinical Pediatrics and Neurology, The Ohio State University College of Medicine, Columbus, Ohio
⁴Department of Pediatrics, Weill Cornell Medical College, New York, NY
⁵Komansky Center for Children’s Health, NewYork-Presbyterian Hospital, New York, NY

Background: Using electronic health information to predict health outcomes and identify high-risk patients requires mapping coded data to clinical concepts. In pediatric medicine, published algorithms can accurately map diagnosis codes (ICD-9) to clinical concepts such as “chronic illness” and “medical complexity” (Simon et al, 2014). However, little is known about how to map coded data to the concept of “technology dependence”. Technology dependent children are a medically fragile group who require (1) a medical device to support a vital body function and (2) intensive ongoing nursing care (Haffner et al, 2001). Identifying technology dependent children is important because they often benefit from improved care coordination. Here, we describe preliminary work to develop and validate an operational definition of technology dependence based on administrative data. This definition will be useful to identify children for care management and to generate variables (features) for use in predictive models. We analyzed claims data from 2011-2012 from Partners for Kids (Pediatric Accountable Care Organization for a group of children followed in a pediatric epilepsy clinic (Nationwide Children’s Hospital, Columbus, OH).

Methods: We chose four pieces of technology for our preliminary analysis: wheelchair, tracheostomy, gastrostomy tube (G-Tube), and ventriculoperitoneal shunting (VP Shunt). We identified codes for these procedures using the Current Procedural Terminology (CPT) 2012 edition and the International Classification of Diseases, Ninth Revision (ICD-9). We generated a list of patients with claims containing these codes. We then manually reviewed available patient charts to determine the validity of the codes. We created Venn diagrams to illustrate the number of patients found with (A) only ICD-9 codes, (B) only CPT codes, and (C) both. We calculated positive predictive value (precision) for each set of codes.

Results: In a group of 1996 children with epilepsy, we found 46 patients with codes for wheelchair, 17 for tracheostomy, 113 for G-tube, and 74 for VP shunt. For wheelchair and G-tube, CPT and ICD-9 codes were both helpful in finding patients. For example, for wheelchair, the CPT codes 97542, 97750, and 97760 correctly identified 15 of 39 patients, and the ICD-9 code of V46.3 correctly identified 3 of 4 patients. For Tracheostomy and VP Shunt, the CPT codes did not identify any patients beyond those found with ICD-9 codes. After excluding patients without charts available for review, the positive predictive values (precision) for the codes were as follows: wheelchair 42% (18 of 43), tracheostomy 86% (12 of 14), G-tube 86% (73 of 85), VP shunt 86% (62 of 72). The figures illustrate the codes used, the number of patients identified with each set of codes, and the number of patients without charts.

Conclusion: Administrative data can identify children with technology dependence with high positive predictive value (precision) for some technologies (tracheostomy, G-tube, and VP-shunt), but not for others (wheelchair). Use of both the CPT codes and ICD-9 codes generated a more complete list of patients than either set of codes alone.

Future Directions: We will continue to refine our definition of technology dependence. We hope to gain access to additional datasets to validate our definitions. Eventually, we would like to expand to other examples of technology dependence (dialysis, implanted insulin pump, implanted baclofen pump, etc.).
Understanding Ongoing Concerns after Implementation of Patient-Provider Messaging in the Acute Care Setting

John R. Hanna¹, Kelly McNally¹, Sucheta Ravindran MS¹ Diana Stade¹, Patricia C. Dykes, RN, PhD¹², and David W. Bates¹², MD, MSc, and Anuj K. Dalal, MD¹²
¹Brigham and Women’s Hospital, Boston, MA, ²Harvard Medical School, Boston, MA

Introduction: Bedside patient portals that incorporate patient-provider messaging functionality have much potential in improving overall communication in the acute care setting.¹ However, providers have expressed concerns about managing expectations for patient use of messaging portals. We previously identified barriers to participation by physicians.² Many were concerned with impact on clinical workflow, decreased patient interaction, and malpractice risk.² The purpose of this study is to understand ongoing concerns regarding patient–provider messaging in the context of actual patient-provider messaging data after implementing a patient portal in the acute care setting.

Methods: We measured patient-provider messaging activity on intervention units (MICU and Oncology) over a two-month period (1/1/15-3/1/15). The messaging tool was contained within a patient-centered toolkit (on a hospital-issued password-protected iPad) available at the bedside to patients and caregivers, which allowed for instant messaging to members of the care team that were on service. Care team members would receive an email notification informing them that a patient had sent them a message. We analyzed content of patient messages and grouped by major themes using a 2-person consensus approach. We conducted a 1-hour focus group consisting of 10 general medicine residents who had worked on the intervention units. During the session we reviewed intended uses of patient-provider messaging and engaged participants in discussing ongoing concerns. We then reviewed actual usage data and a preliminary analysis of patient message content with participants.

Results: Of the 231 patients/caregivers who used the toolkit, one hundred and fifty-eight sent at least one message to their providers. Thirty-six messages per week or 146 messages per month were sent to providers over the two-month period. In total, patients/caregivers sent an average of 2 messages per hospitalization. Our analysis of message content yielded the following: 19% contained health concerns, needs, and questions, 11% reported clinical status or symptoms, 11% pertained to test results or medications, 6% involved care coordination (i.e. “ICP will arrive at 12:30 this afternoon”), 26% was feedback for the care team and 42% were study related or gibberish (i.e. “I am an iPAd User”). We identified three main areas of concerns expressed by providers: 1) use for conveying emergent clinical signs and symptoms; 2) excessive patient messaging with the potential for workflow disruption due to a rapid influx of notifications; and 3) reduced face-to-face interaction with the potential for adversely impacting patient experience and quality of care.

Discussion: We observed that the volume of messages sent by patients was low and patients mostly used the messaging feature to provide non-urgent input and feedback to providers. Ongoing provider concerns regarding inappropriate and excessive use were not consistent with actual message content or messaging activity observed. Furthermore, the relative dearth of messages sent by patients argues against concerns of reduced face-to-face interactions between physicians and patients. Despite presenting these findings to providers who used the messaging tool, ongoing concerns regarding future use by patients lingered. Managing expectations regarding appropriate use of portals for messaging is essential, and could lead to further uptake by providers and ultimately increase patient utilization. An effective messaging tool could supplement and reduce the use of pagers and email in the clinical setting. By placing all members of the care team in one “chat room” and allowing clinicians to “choose” who gets notified of a message, clinicians can reduce the amount of unnecessary emails and pages and focus on improving quality of care.

Acknowledgements: The Brigham and Women’s Hospital PROSPECT project is part of the Libretto Consortium supported by the Gordon and Betty Moore Foundation

References
Using a Socio-Technical Framework to Understand Technology Use Among Health Care Innovation Award Community Resource Planning Awardees

Saira N. Haque, PhD, MHSA, Lucia Rojas-Smith, DrPH, Debra J. Holden, PhD
RTI International, Research Triangle Park, NC

Abstract
The Centers for Medicare and Medicaid Services is supporting a number of innovations to improve community based outcomes through the Health Care Innovation Awards (HCIA). We conducted an evaluation of a subset of HCIA awardees with a focus on community resource planning to understand the impact of the innovations. A qualitative analysis was conducted using a sociotechnical framework for awardees with informatics components as part of the innovation. While the informatics components varied across awardees, studying them using a sociotechnical framework assisted in identifying themes about technology use and integration in the innovation as a whole.

Introduction
Community-based health interventions have potential to improve health outcomes of community members. In order to support innovations in service delivery or payment to improve outcomes at the community level, the Center for Medicare and Medicaid Innovation within the Centers for Medicare and Medicaid Services (CMS) funded the Health Care Innovation Awards (HCIA) Program [1]. The Community Resource Planning (CRP) programs are designed to improve health, health care and costs of care, particularly for Medicare, Medicaid and Children’s Health Insurance beneficiaries. HCIA CRP awardees all incorporate technology into their innovations to varying degrees. Eight of them have informatics as a primary focus. As part of a larger evaluation currently in progress, we sought to more fully understand how the HCIA CRP awardees with a focus on informatics (n=8) are using innovative methods to integrate people, processes, technology and context to improve community outcomes.

Methods
A qualitative analysis was conducted for the 8 awardees with an informatics focus. Data collection methods included site visits, qualitative interviews and document review. Document review was conducted of the initial grant application to the program, quarterly reporting by awardees, and documents provided by the awardee. A semi-structured interview guide was developed and tailored by role for technical, clinical and administrative staff at each awardee. Interviewees were identified in concert with awardees. The interviews were audiorecorded and notes taken. Documents were uploaded into NVIVO for coding. Thematic analysis was conducted using an interview guide based on the sociotechnical framework [2].

Results
There were a wide range of ways in which informatics was integrated into the Community Resource Planning Awardee efforts. These include support for information availability for patients across organizations, alerts for patients throughout a community, facilitating coordination, management of chronic illness, and reporting based on health information exchange. Informatics innovations were also used to support stratification of patients into risk tiers so that the work of care coordinators could be appropriately directed. Despite the wide range of informatics innovations, where were a number of common themes in its use. The degree to which the technology was aligned with organizational and community structures varied. In addition, technology and workflow integration differed among awardees. Factors that influenced informatics integration into the overall innovation included the strength of partnerships, fit between workflow and technology, technology understanding and acceptance, knowledge of the innovation by clinicians and education.

Conclusion
While evaluation work is still in progress, these activities have great promise to improve health, improve care and decrease costs. The HCIAs showcase emerging techniques for improving health through community partnerships. Findings will help communities better understand which innovative technology efforts have promise for improving community outcomes and adapt them based on their own community needs and technological capacity.

References
Standardized Mapping of Sensitive Data Categories  
Tiffany L. Harman RN, Rachael A. Howe RN, BSN  
3M Health Information Systems, Inc., Murray, Utah

Introduction
The widespread adoption and use of digital health information exchange between healthcare organizations to support the flow of health information to where and when it is needed is undoubtedly crucial. However, concerns arise as to what data can be shared and what data must be withheld if it is sensitive health information. Sensitive data categorization can be determined by the relevant federal and state laws, and can differ to a certain degree by state, system or use case. For semantic interoperability, data being exchanged should be encoded using standard terminologies. If the members of the sensitive data categories have been encoded using standard terminologies, then these standard codes can be automatically flagged as not to transmit (or accept). In order to evaluate whether standard terminologies provide sufficient coverage, we mapped the concept members of the following sensitive data categories: AIDS/HIV, Developmental Disability/Mental Health, and Genetic Testing. Because no one standard terminology contains all the data elements needed for sensitive health information, multiple terminologies were leveraged for this mapping. To start with, Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) was used to encode diagnosis and Logical Observation Identifiers Names and Codes (LOINC) was used to encode lab test and results. However, mapping to reimbursement terminologies such as the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM), ICD-10-CM, ICD-10 Procedure Coding System (ICD-10-PCS), Current Procedural Terminology (CPT), and Medicare Severity Diagnosis Related Group (MS DRG) was also performed, as data could also be encoded with these standards. Medications were not included in this exercise.

Method
The project had four phases: 1) evaluate each sensitive data category to identify member concepts; 2) identify the standard terminologies applicable to the concepts (based on the concept definition); 3) map each member concept to one or more standard terminology codes, and 4) implement the mapping in a terminology server to provide a practical maintenance solution.

Results
6771 standard codes were mapped to the members of the sensitive data categories listed above. Of these, 40% of the coverage was represented by SNOMED CT and 32% of the coverage was represented by LOINC. Note that a concept can be a member of more than one sensitive data category. There are 34 unique standard codes that are in different sensitive data categories. There are 2 SNOMED CT codes and 2 LOINC codes that are in both the AIDS/HIV and Developmental Disability/Mental Health sensitive data categories, and there are 31 SNOMED CT codes and 31 LOINC codes that are in both the Developmental Disability/Mental Health and Genetic Testing sensitive data categories (see figure 1).

Conclusions
Identifying standard codes that correspond to sensitive data can help to set up automated flagging to prevent unauthorized distribution. However, there are implementation challenges. There is no standardized list of concepts and standard codes that constitute sensitive data, resulting in variability in determining which data should not be shared. Even with consensus, there could be false negatives. For example, patient data may have the wrong standard code assigned in the electronic health record (EHR), or the EHR may not have a standard code at all. Resolving data issues requires collaboration to ensure data integrity, proper use case and continuous improvement. Nevertheless, creating a custom categorization and encoding sensitive health information to standard terminologies is a useful and usable first step.
Before-after implementation of the sniffer for the detection of failure to recognize and treat severe sepsis

Harrison AM, BS; Thongprayoon, C, MD; Park JG, MD; Daniels CE, MD; Clements CM, MD, PhD; Goyal DG, MD; Elmer JL, DNP, APRN, CNS; Gajic O, MD, MSc; Pickering BW, MB, BCh, MSc; Herasevich V, MD, PhD
Mayo Clinic, Rochester, MN

Background: In 2013, the US Agency for Healthcare Quality and Research ranked septicemia as the most expensive in-hospital condition in the US, based on 2011 data.1 Despite increased understanding of this problem,2 previous sepsis detection and alert systems have failed to demonstrate improvement in clinically meaningful outcomes.3 The objective of this study was to measure the impact of the automated, electronic sepsis sniffer on compliance with the 3-hour Surviving Sepsis Campaign (SSC) bundle elements.

Methods: We recently developed and validated a sniffer for the detection of failure to recognize and treat severe sepsis.4 Beginning 01/14/2015, the sepsis sniffer (suspicion of infection, Systemic Inflammatory Response Syndrome, and organ dysfunction components) was implemented in the medical ICU (MICU) at Mayo Clinic Rochester in AWARE—a patient-viewer and clinical decision support tool designed at Mayo Clinic to reduce risk of error and already in routine use in the MICU5—as well as in the Emergency Department (ED). Implementation occurred after focused provider training and presentations. Using METRIC Data Mart,6 a relational database and near-real time duplicate of the complete hospital EMR at Mayo Clinic, data on 3-hour SSC bundle element compliance for patients before implementation of the sepsis sniffer in the MICU (January through March 2013, N=98) was compared to MICU and ED data after implementation of the sepsis sniffer (January 14th through March 2015, N=60). As some ED visits result in no hospital admission or no admission to the MICU, these outcomes are not present for all patients in the “after” cohort. Average time to completion was calculated as: “Time of bundle element completion” minus “Time of sepsis detection”. Thus, negative results indicate compliance before detection.

Results: Overall percent completion of the 3-hour SSC bundle was increased after tool implementation. Overall average time to bundle completion was reduced for all 4 elements. Detailed results are in the following Table.

<table>
<thead>
<tr>
<th>Percent completion (%)</th>
<th>BEFORE (N = 98)</th>
<th>AFTER (N = 60)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-hour SSC Bundle overall compliance</td>
<td>25</td>
<td>55</td>
<td>0.001</td>
</tr>
<tr>
<td>Measure lactate level</td>
<td>64</td>
<td>95</td>
<td>0.001</td>
</tr>
<tr>
<td>Obtain blood cultures</td>
<td>81</td>
<td>88</td>
<td>0.204</td>
</tr>
<tr>
<td>Administer antibiotics</td>
<td>61</td>
<td>93</td>
<td>0.001</td>
</tr>
<tr>
<td>Administer 30 mL/kg fluids</td>
<td>46</td>
<td>62</td>
<td>0.055</td>
</tr>
<tr>
<td>Average time to completion median value, mins (IRQ)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-hour SSC Bundle overall compliance</td>
<td>-4 (-137 to 148)</td>
<td>-126 (-159 to -18)</td>
<td>0.184</td>
</tr>
<tr>
<td>Measure lactate level</td>
<td>128 (0 to 374)</td>
<td>13 (-63 to 125)</td>
<td>0.002</td>
</tr>
<tr>
<td>Obtain blood cultures</td>
<td>416 (79 to 710)</td>
<td>153 (53 to 517)</td>
<td>0.019</td>
</tr>
<tr>
<td>Administer antibiotics</td>
<td>116 (8 to 398)</td>
<td>-31 (-84 to 85)</td>
<td>0.001</td>
</tr>
<tr>
<td>Administer 30 mL/kg fluids</td>
<td>-4 (-138 to 332)</td>
<td>-120 (-159 to 105)</td>
<td>0.380</td>
</tr>
</tbody>
</table>

Conclusions: Implementation of the sepsis sniffer improved overall 3-hour SSC bundle compliance with respect to percent completion, as well as average time to completion for some elements.

References:
Meal-Camera System to Support Nutritional Tele-Consultation for Diabetics

Masayuki Hashimoto1, MSc, PhD, Umi Nakamura1, Miki Akiyama, MSc, PhD2, Toru Takebayashi, MD, MPH, DMSc2, Marino Oshima3, Tomoki Kumegawa3, Hitomi Imachi, MD, PhD 3, Koji Murao, MD, PhD3

1KDDI R&D Labs., Saitama, Japan; 2Keio University, Tokyo/Kanagawa, Japan; 3Kagawa University Graduate School of Medicine, Kagawa, Japan

Abstract
We propose a communication tool for connecting diabetic outpatients and hospital’s medical staff such as dietitians and doctors. This system supports nutritional tele-consultation based on meal images taken by patient’s smartphone. Patients easily send images of their meals with applications installed in smartphones. Dietitians use computers and guide the patients remotely based on the images. Clinical study with 20 patients show that patients’ weight and HbA1c improved in the most cases.

Introduction
In order to prevent deterioration in diabetic’s condition, it is essential to guide their nutrition consumption every day. It is difficult for hospital medical staff to observe outpatients’ daily diet, thus some patients often do not follow nutritional guidance that they should follow. This causes serious deterioration in their condition. Therefore, we developed “Meal-camera system” to support daily nutritional tele-consultation. Its smartphone app for patients is easy to use and its web interface for dietitians is designed for supporting quick nutritional evaluation of patient’s meal. Nutritional tele-consultation is effective for motivation to continue health program[1]. Here we show the outcome in clinical study with diabetic outpatients.

Methods and Results
The system consists of a smartphone with an original application, an image-storage database server and PC for dietitians in a hospital. Patients take pictures of all meals (including snacks) and send images using smartphones. Hospital dietitians reply comments once or twice a day using PC. Dietitians and patients can communicate with each other using SMS function in this system. In our clinical study, 20 diabetic patients received this nutritional tele-consultation. The patients’ weight and HbA1c improved in the most cases. Followings are details of two cases of them:

[Case1] A pregnant woman with diabetes type 1, who miscarried previously due to hyperglycemia, became pregnant. Too much fear on hyperglycemia leads her to inject insulin excessively. That caused failure in controlling her blood sugar. Thus we started tele-consultation using the proposed system. After that, her blood sugar became stably controlled. She gave birth without any problems at last.

[Case2] A woman with diabetes type 2 took medicine in obesity treatment, but it was not effective. Furthermore, it was difficult for dietitians to assess her daily intake by interviews. Accordingly, we started nutritional guidance using the system. After that, the patient worked on dieting. This caused improvements in both losing weight and HbA1c shown in Figures 1 and 2.

Conclusion
We developed “meal-camera system” to support daily nutritional tele-consultation. In clinical study, 20 diabetic outpatients consulted hospital dietitians. The results show effectiveness of the tele-consultation using this system.

References
Conducting health insurance surveillance with electronic health records

Brigit Hatch, MD MPH1; Carrie Tillotson, MPH1,2; Megan Hoopes, MPH2; Miguel Marino1, PhD; Rachel Gold, PhD MPH3; Jennifer E. DeVoe, MD DPhil1,2

(1) Oregon Health & Science University, Portland, Oregon, USA; (2) OCHIN, Inc, Portland, Oregon, USA; (3) Kaiser Permanente Northwest Center for Health Research, Portland, Oregon, USA

Purpose: Under the Affordable Care Act (ACA), Americans have new access to health insurance, yet challenges remain to successfully obtaining and maintaining coverage. The Electronic Health Record (EHR) presents new opportunities for patient-centered primary care homes to track patients’ insurance coverage and assist with enrollment and retention. We aimed to utilize EHR data from a national network of community health centers to monitor insurance status over time and identify factors associated with uninsurance.

Methods: Using EHR data from the OCHIN network of over 300 clinics in 16 states, we conducted a retrospective cohort study of all adults seen for primary care at an affiliated clinic from 2012-2013. We described the population and its utilization of primary care services, monitored changes in clinic-identified insurance status over time, and used generalized estimating equation (GEE) logistic regression with a robust sandwich estimator to identify factors associated with uninsurance.

Results: The population (n=279,654) was predominantly female (58%), white (77%), non-Hispanic (71%), and living with incomes beneath the federal poverty line (57%). At their first visit during the study period, the majority (41%) were uninsured, 39% had Medicaid, 16% had commercial insurance, and 4% had other insurance types. Among patients with no insurance at their first visit within the study period (n=114,000), 50% continued to be uninsured for every subsequent visit during the study period, while 13% gained insurance, and 36% had no further visits during the study period and therefore unknown insurance status. Among patients with Medicaid at their first visit (n=109,431), 68% maintained insurance at every subsequent visit, while 7% had lost insurance, and 24% had no subsequent visits. Among patients with commercial insurance at their first visit (n=44,420), 63% maintained insurance, 6% lost insurance, and 31% had no further visits. Compared to patients with Medicaid, factors associated higher odds of uninsurance included female gender, age >25, Hispanic ethnicity, income above the FPL, and rural residence (p<0.01).

Conclusions: This study identifies a large group of patients who continue to seek care at community health centers despite ongoing lack of health insurance and confirms previous studies highlighting the social vulnerability of this population. Now that many of these patients can access health insurance under the ACA, there is critical need to develop systems to support patients in obtaining and maintaining coverage. The electronic health record presents new opportunities for identifying and tracking insurance status within a patient-centered primary care home. Currently, the EHR can only track insurance status in relation to each office visit. However, EHR-based tracking tools could capture automated data and support health insurance outreach from within the patient-centered primary care home both during and between patient visits. As the delivery system moves toward care beyond office visits and as more people than ever have access to public insurance, such methods hold promise for assisting vulnerable populations with the challenge of obtaining and maintaining health insurance.
Mapping Hospital Infections to Inform Quality Improvement Interventions

Courtney L. Hebert, MD, MS¹,², Tara B. Borlawsky, MA¹
¹The Ohio State University, Department of Biomedical Informatics, Columbus, OH; ²The Ohio State University, Division of Infectious Diseases, Columbus, OH

Abstract
Hospital acquired infections (HAIs) are the target of federal initiatives to improve quality of care. HAI surveillance is often performed manually, and reported using tables and charts. We assessed the feasibility of automating an HAI identification algorithm, and displaying the results on a hospital map to provide geospatial context.

Introduction
Reducing hospital-acquired infections (HAIs) has become a leading focus of quality improvement efforts nationwide. Despite the growth of electronic health records (EHRs), surveillance data used to inform these efforts are often manually abstracted. This takes significant time and effort on the part of safety and quality staff, and results in outdated data. Public health surveillance data are often displayed on maps in order to understand the potential transmission and outbreak pattern of a public health concern, however, when hospital surveillance data are shared between safety and quality staff they are most commonly presented in tabular format or line graphs. In fact, for many infectious diseases the geospatial distribution can be extremely important for evaluating transmission between patients or for potentially identifying unit-related differences in care-processes. As part of an ongoing effort, we propose to create a hotspot map of the hospital, displaying the location of each HAI, in order to identify high risk areas and inform targeted interventions to decrease HAIs (i.e. unit specific training).

Methods
The main objectives of this project are to: (1) create an algorithm for normalizing and standardizing the organisms to which a catheter-associated urinary tract infection (CAUTI) is attributed, and (2) display the location of these events on a map of the hospital. Data on all hospital-acquired CAUTIs reported per the Centers for Disease Control National Healthcare Safety Network definition¹ at The Ohio State University Wexner Medical Center over a 1-year period were collected along with clinical data including microbiology and locations. To develop an automated microbiology algorithm we also collected 3 months of general microbiology results. These were parsed, and annotated with SNOMED-CT codes to allow us to systematically identify organism type, including genus, species, Gram stain and colony counts. The location of attribution of each CAUTI was displayed on a map of the hospital unit as a dot, color-coded to indicate type of organism in order to additionally track potential organism transmission. The dot appears on the map at the date of onset, and is removed on the date of transfer out of the unit.

Results
Of >50,000 microbiology results, 1195 were identified as positive. Using a Perl script to implement a set of regular expression parsing and lexicon matching rules, SNOMED-CT codes for organism and genus were automatically assigned to 1143 organisms (983 bacterial and 160 fungal). While the algorithm has yet to be formally evaluated, an informal assessment by an infectious disease clinician (CH) indicates that it is highly accurate. The microbiology parsing algorithm was then applied to the CAUTI data. The location of each CAUTI was displayed on a map of the hospital, labeled with infection type. The CAUTIs were not uniform across hospital units, but clustered in units with higher risk patients. The distribution of organisms and temporal association between CAUTI events were easily visualized.

Discussion
In this study we were able to successfully parse microbiology reports and show that a map of surveillance data provides useful context, not currently available in tabular reports. Future work includes a formal evaluation of the microbiology parsing algorithm, development of an automated CAUTI identification algorithm and creation of a real-time, interactive map of the hospital displaying multiple types of HAIs and infecting organisms over time, which could be used by infection preventionists to evaluate potential outbreaks and inform their interventions.

References

1495
Identifying ECG Features in Congenital Heart Disease Using Variants of Dynamic Time Warping

Emily P. Hendryx, M.A. 1, Craig G. Rusin, Ph.D. 2, Beatrice M. Riviere, Ph.D. 1
1Rice University, Houston, Texas; 2Baylor College of Medicine, Houston, Texas

In the clinical setting, bedside monitors provide physicians with large amounts of physiological data. One of the signals presented on the screen is the electrocardiogram, or ECG. Each feature of the ECG corresponds to a specific part of the cardiac cycle, and to the well trained eye, an ECG can provide key information about a patient’s clinical status. Because of their diagnostic utility, the primary goal of this research is the automatic identification of ECG features in congenital heart disease. These features can then in turn be used to develop real-time predictive models to help doctors make informed treatment decisions.

With data from Texas Children’s Hospital, our work focuses on physiological signals from patients with hypoplastic left heart syndrome (HLHS). Patients with this congenital heart disease have an underdeveloped left ventricle, resulting in a lack of blood flow to the rest of the body. Because these are sick patients, their ECG signals do not look like the typical textbook ECGs and seem to vary widely. Adding more complexity, data is usually simultaneously recorded from four ECG leads, each lead providing information from a different angle. Our goal is to effectively sort the large amounts of ECG data available into classes based on heartbeat morphology and then identify ECG features based on their classification.

To approach this problem, we first combine the different ECG lead information into one signal, the vectorcardiogram magnitude, reducing the number of individual beats to be analyzed. In clustering the data, variations of dynamic time warping are used in conjunction with k-medoid clustering. Dynamic time warping (originally used in speech recognition) allows for the comparison of two signals that might have similar features, but these features are shifted in time. Hence, it is a measure that lends itself readily to ECG feature identification. The methodology is tested on a set of synthetic ECG data and on a set of approximately 4000 patient heartbeats with promising preliminary results.

With the identification of the representative morphology classes found in the data, dynamic time warping then provides a means for mapping known features in a pre-labeled heart beat to similar un-labeled heart beats. While the application of dynamic time warping to ECG data is not new, there appears to be little literature based on pediatric data – especially data from patients with congenital heart disease. This work addresses this gap in the literature and provides a stepping stone for the development of ECG-based predictive models for the pediatric population.

This research was funded by a training fellowship from the Keck Center of the Gulf Coast Consortia, on the Training Program in Biomedical Informatics, National Library of Medicine (NLM) T15LM007093, 2014-2015.

References

Using MetaMap to Analyze Which Linguistic Concepts in an Imaging Study Indication Make it Helpful to a Radiologist

Andrew Herzik, MD1, Joshua M. Pevnick, MD, MSHS2
1LAC+USC, Los Angeles, CA; 2Cedars-Sinai, Los Angeles, CA

Abstract
For reading radiologists, order indications improve interpretation. We used UMLS MetaMap to analyze which linguistic concepts in order indications are most helpful to radiologists. We analyzed indications from 200 abdomen/pelvis CTs, and found that the following concepts were most associated with higher indication quality: “Modifying concepts” (p=0.01), “Findings” (p <0.0001), “Disease/Syndromes” (p=0.01), and “Ideas or Concepts” (p = 0.01). This knowledge could be used to help improve providers’ indication quality.

Background: Order indications for radiology exams provide valuable clinical context to reading radiologists; this, in turn, facilitates more successful interpretation. Unfortunately, there is substantial variation in the quality of radiological order indications. We are aware of studies analyzing how structural characteristics, like computerized physician order entry (CPOE), affect indication quality. However, we sought to learn which linguistic concepts in the indication would be most helpful to a radiologist, and were unaware of any prior research in this area.

Methods: We conducted an IRB-approved, retrospective analysis of de-identified radiological order indications from a large university-affiliated hospital. We obtained order indications from 200 randomly selected inpatient CT abdomen/pelvis studies ordered between January 1st, 2012 and June 30th, 2012. 100 indications were from a pre-CPOE environment and were handwritten on paper that was otherwise blank. The remaining 100 were ordered with CPOE that included study-specific buttons and the option of including unstructured text.

To extract linguistic concepts, we used the text mining UMLS Metathesaurus MetaMap to map order indications. We compared MetaMap results to two radiologists’ independent implicit Likert scale ratings, which we used as a reference standard scale of indication quality. Two radiologists with extensive experience reading CT scans of the abdomen/pelvis independently rated each order indication on a 5 point Likert scale (“The following indication helps me successfully interpret the CT abdomen study”). We hypothesized that MetaMap categories “location concepts” and “temporal concepts” would be most predictive of high scoring indications.

Each order indication was mapped using MetaMap 2014 on default settings (data source: 14/15 Transition – 2014AB, data version: USAbase, data model: strict). In the instance of multiple concepts, we used clinical judgment to determine the predominant concept. We also used clinical judgment to overrule incorrectly identified concepts (e.g. spelling errors erroneously identified as concepts), which were then discarded. We used linear regression to examine the effect of various language concepts on the reference standard score.

Results: We analyzed 200 CT abdomen/pelvis order indications. We generated a final model using stepwise regression. In our final model, use of “Temporal concepts” (p=0.025), “Modifying concepts” (p=0.01), “Signs and Symptoms” (p=0.032), “Findings” (p<0.0001), “Disease/Syndromes” (p=0.014), and “Ideas or Concepts” (p=0.01) were significantly associated with higher Global Implicit Scores, (model R²=0.33).

Discussion: Understanding which linguistic concepts are most helpful could be used by ordering clinicians to improve indication quality. Furthermore, it could be used by designers of electronic health records to choose which text to offer on study-specific indication buttons. Simple modifying concepts such as “severe” or “worsening” for example, if attached to basic findings would add over half a point in the quality of indication. Electronic health records could potentially provide automated real-time feedback to ordering clinicians regarding their indication quality. Additional research is required to understand how variables like CPOE implementation or physician specialty shape linguistic patterns. Nonetheless, this study demonstrates the potential utility of data mapping with MetaMap to inform and improve communication between physicians.

Electronic Health Record Audit Logs: An Alternative Approach to Workflow Analysis

Authors
Annemarie G. Hirsch, PhD, MPH1, Virginia R. Lerch, MPH2, J.B. Jones, PhD. MBA3, Walter F. Stewart, PhD, MPH1
1. Center for Health Research, Geisinger Health System, Danville, PA; 2. Institute for Advanced Application, Geisinger Health System, Danville, PA; 3. Research Development and Dissemination, Sutter Health, Walnut Creek, CA

Background
As described in Sittig and Singh’s socio-technical model for studying health information technology, successful development, implementation, and evaluation of health information technology interventions in clinical settings are dependent upon the ability to assess and adapt to clinical workflows.1 Consistent with this model, studies of electronic health record adoption have revealed the barriers presented by incompatible workflows and the success associated with thorough workflow analysis prior to HIT implementation. The pace at which new HIT is becoming available in clinical care settings is increasing. The HITECH Act has triggered a rapid increase in EHR adoption and now EHR vendors such as EPIC are beginning to make EHR integrated apps available to health systems. Traditional methods of workflow analysis, such as time and motion studies, will be unable to keep pace with what is anticipated to be a rush of new technology in clinical care settings, putting at risk the ability to successfully integrate these potentially beneficial technologies while also hindering assessment of the impact of these technologies on clinical care.

The EHR audit file, a log that tracks user activity in the EHR, is largely an untapped source of data that provides an alternative to resource and time intensive methods of workflow analysis. The EHR audit log file maintains a time-stamped audit log of transactions when a patient record is in use. The audit log tracks users (i.e., nurses and providers who log in to patient record), actions (i.e., “view”, “accept”, “cancel”, “print” and “exit”), and the EHR modules in which the actions take place (e.g., encounter, medication lists, orders, etc.). Previous efforts have used audit files to measure clinical documentation time, detect patterns of EHR use, and monitor user activity—we describe another method of utilizing the audit file to understand clinical workflows.

Methods
We used the audit file data from the Geisinger Health System (GHS) EHR to better understand the workflow of primary care visits in family practice clinics from the time the patient checks in to when the patient checks-out. Specifically, we used retrospective audit file data to determine what health care providers are involved in primary care visits; in what order the providers work with patients; how much time providers spend with the patients; and how much time patients wait during a primary care encounter.

Results
Encounters were included in the analysis if they were for primary care patients; took place in one of the 30 GHS family practice clinics; had at least one action recorded in the audit file between check-in and check-out; and took place between January 1, 2009 and June 30, 2011. There were 1,212,446 encounters in the 30 GHS family practice clinics between January 1, 2009 and June 30, 2011. To reduce the processing time and computational memory required to analyze more than 1 million encounters, we randomly selected 30 dates on which at least one of these encounters occurred using simple random sampling. Applying the eligibility criteria to the encounters scheduled on these 30 dates reduced the number of encounters to 36,437 across 26 separate family practice clinics, 3.2% of all family practice encounters during the study period.

On average, patients spent 5 minutes with the nurse and 16 minutes with the provider during these encounters. Patients spent an average of 14 minutes in the waiting room, 11.6 minutes in the exam room waiting for the provider, and 9.7 minutes waiting to check-out of the encounter.

Conclusion
Our data confirm the findings of prior work that used survey and direct observation to measure duration of primary care visits.2 The ability to study the audit data retrospectively makes this low cost method of workflow analysis a viable alternative to the use of more traditional methods of workflow analysis, such as time and motion studies, which require a prospective study design and need advanced notice of pending policy changes in order to evaluate the impact of these changes. The longitudinal data available in the audit files can also be linked to encounter-level and patient-level clinical data to offer unique insights on the impact of the frequent and rapid changes in operations and health policy.

References
‘Smart Snack Box’ System for Recording Snacking Behavior

Teruyoshi Hishiki, MD, PhD¹, Hiroki Yasui, MD, PhD², Yoshiko Matsunaga, RN, PhD¹, Takuro Tamura³

¹Toho University, Tokyo, Japan; ²Nagoya University, Nagoya, Japan; ³LINE Co., Ltd, Tokyo, Japan

Abstract
A convenient method of identifying individual snacking patterns would be helpful in dietary consultations. Here, we report an ongoing project to develop a ‘Smart Snack Box’ that records the timing of snacking behavior. The apparatus contains sufficient snacks for an individual to eat during the day, and it electronically detects and records each opening of the lid. We plan to test the system on healthy subjects.

Introduction
Studies of the circadian rhythm in humans have revealed the impact of the frequency and timing of meals on the metabolic state, which indicates that eating patterns have an influence on health and disease¹. The effects of snacking patterns on body weight and diabetes control have also been investigated². To apply this knowledge, it would be useful to profile the frequency and timing of snacking in individual subjects. Here, we report the ongoing development of a ‘Smart Snack Box’ that records the timing of snacking behavior.

Methods
The box consists of a container with a hinged lid, a microcomputer, and a controller computer. A Hall effect sensor is positioned on the side wall and connected to the microcomputer, and a magnet is positioned on the lid such that it almost touches the sensor when the lid is closed. The microcomputer sends an ‘OFF’ signal to the controller computer while the lid is open. While the lid is closed, the sensor generates a voltage and the microcomputer sends an ‘ON’ signal to the controller computer. The accuracy of the sensor was checked while opening and closing the lid 100 times prior to the full experiment.

The setting of the experiment is as follows. The intended subjects are three male and three female office workers to be recruited by the authors. We will record their snacking behavior, using their declared snack food, over three working days. Before the system is activated, the user will fill the box with sufficient snacks for consumption during the working day. The user is free to open the box and remove a snack whenever he/she likes. Timestamps of the snacking behavior will be logged on the computer, and snacking patterns over the course of the day will be extracted for each subject. We will discuss the results of the analysis with the participants, focusing on whether they were aware of their eating patterns and why they eat snacks at certain times of the day.

Results and Discussion
The sensor correctly detected the opening and closing of the lid for each of the 100 repetitions. There are few methods for collecting data on eating behavior that do not rely on self-reporting, except for a technique for detecting chewing. A limitation of the proposed method is its inability to record the quantity of snacks removed. This could be overcome by integrating a scale into the snack box. Collection of individual snacking patterns, as well as recording of physiological information such as heart rate or blood pressure, would enable the ‘Smart Snack Box’ to control snacking behavior by directing the user to change their behavior at the time they are likely to reach for a snack.

Conclusion
We have developed the ‘Smart Snack Box’ as a method of recording snacking behavior without self-reporting. The next step of testing is a trial to be conducted with healthy volunteers.

References
Introduction: Clinical practice guidelines have a slow rate of adoption in medicine. Clinicians need timely and widespread dissemination of evidence to guide clinical practice. Really simple syndication (RSS, also called rich site summary), a method ubiquitous to news outlets, could also be used to circulate journal articles.

Methods: In a systematic review, we examined the use of RSS by medical professionals. We reviewed several databases (LISTA, Pubmed, and Ovid(Medline/CINAHL/Psych/Health&Bus) with search inclusion criteria of the following keywords: “RSS, rich site summary, really simple syndication” as well as MeSH term Webcasts (where RSS topic is found). An article was excluded if it was not written in English, if the paper did not cover RSS feeds, or if the paper was not about Medicine.

Results: Our review initially found 1610 articles, of which 1430 were excluded as duplicates, irrelevant to the topic or by abstract review. In full text screen of the remaining 183 articles, 50 more articles were excluded for various reasons. For this review, we included 130 articles, and we found several categories of use within medicine, like nursing, veterinary medicine and health librarian education. A few articles explained RSS and its benefits to medical professionals. Of the articles screened, many were published by non-medical sources (i.e., Fortune or Forbes magazine) or were opinion pieces in specialty journals lacking significant evidence or rigor.

Conclusion: We found a number of different articles with explanations of how medical professionals use RSS. Most of these are found without descriptions on how to set up a personalized feed, and are more anecdotal in nature. We have summarized the guidance of these articles to help clinicians set up RSS feeds. This step by step guide will direct the clinician to different websites based on interests and will help them navigate web pages of journals and information sites to find sources to add to their feeds. Once added, the feeds will become more personalized, and will serve as a comprehensive platform to help clinicians use RSS in effective and relevant ways.

Katherine A. Homann, M.D. 1, Adam C. Rule, M.S.1, Skylar G. Kerzner B.S.1
1University of California San Diego, San Diego, California

Introduction to the Problem and Specific Purpose of Research

Price transparency has been proposed to decrease rising healthcare costs.1 Large interventions like Aetna’s Member Payment Estimator and Castlight’s tools for WellPoint show price transparency to be a growing field in health information technology (HIT).2 However, no systematic review of price transparency existed. Our new systematic review explores how displaying price information impacts the number and cost of services selected by both patients and physicians.

Methodology

Following PRISMA criteria, we searched both PubMed and the Web of Science Core Collection for articles on price transparency, using keywords including “cost,” “charge,” “transparency,” “display,” and “electronic health record.” We limited our search to publications that are written in English, have full text versions available, and contain an original empirical study on patient or provider access to price information.

Results

The initial search yielded 2098 citations, of which 69 met criteria for a full-text review. Ultimately 11 articles met criteria for inclusion in the final analysis. For the eligible publications, study participants were primarily clinicians (9 of 11 studies, or 82%) although two articles described patient-facing price displays. Clinical environments included a range of inpatient, outpatient, and Emergency Department settings in both academic and non-academic centers. Price displays covered a breadth of medical tests including radiology imaging, laboratories, and others.

The two patient studies suggest patients made choices balancing cost and quality, and that price transparency decreased imaging and laboratory costs by a mean of 13%. The nine clinician studies allowed more comprehensive analysis. All clinician studies found a reduction in number of orders, with a mean decrease ranging from 4.5% to 27%. The mean total cost of orders decreased from 3% to 37% with varying statistical significance in those articles; however, all articles measuring cost savings found statistically significant savings in at least one category of medical tests.

Discussion

Nearly half (5 of 11, or 45%) of empiric price transparency articles were published in the last 2 years, indicating the timely relevance of this topic, but also limiting some power of our review. The overall cost savings suggest future research is warranted, and the limitations gleaned are particularly relevant to shape future research. Most studies evaluated the price transparency impact for six months or less, so future longitudinal studies to assess the long term cost impact would be valuable. Also, health outcome measures were assessed in only two of the publications and revealed mixed results, so future research should ensure quality metrics remain high even if costs decrease. Also, due to the range of clinical environments, participants and medical tests studied, further research is indicated—particularly to explore the growing field of patient-centered price transparency.

Conclusion

Price transparency is an HIT tool proposed to lower healthcare costs. Our systematic review reveals cost savings in a range of studies, and suggests further research is warranted to explore this growing field.

References

Informational Content of Verbal Handoffs in Emergency Care

Jan Horsky, PhD,1,2 Edward H. Suh, MD,3,4 Osman Sayan, MD,3,4 Vimla Patel, PhD5

1Brigham and Women’s Hospital, Boston, MA, 2Harvard Medical School, Boston, MA  
3New York-Presbyterian Hospital, NY, 4Columbia University Medical Center, NY,  
5New York Academy of Medicine, NY

Introduction
Handoff of care between shifts is a known risk factor for medical error and adverse events.1 Physicians transferring the care of patients need to succinctly characterize the course of care and convey it to the incoming clinician. Verbal communication is often limited to a brief narrative encompassing laboratory results, significant events, assessments and other often complex data that need to be presented as interconnected events that can be accurately interpreted. Sufficiently detailed presentation will allow the new team to continue care without interruption. Interventions to reduce misinterpretation and loss of facts often involve a combination of training, standardization and the use of structured forms. However, large variations in illness complexity and diagnostic certainty between patients in the same handoff process are intrinsic to emergency care and make uniform and inflexible forms of support unlikely to be effective. Research suggests that handoffs discussions also afford the opportunity to consider alternative viewpoints, correct mistakes and propose new tests or interventions.2 Our objective was to observe discussions during handoffs and analyze how uncertainty affects the content of presentations and to test our assumptions about the relationship between uncertainty levels, clinical complexity and different needs of support for physicians describing patient cases.

Methods
We observed and recorded clinicians during eleven shift changes. Groups of two to three outgoing residents presented between seven and twenty-two patient cases to an incoming attending physician in an ED of a large urban hospital. Recordings containing conversations about 150 patients were transcribed and categorized along dimensions of diagnostic uncertainty and fit with illness script. We operationalized uncertainty along three axes with 3- and 2-point scales. The constructs were derived from concepts in medical education and clinical care and were developed iteratively in a series of meetings and conversations. Patient narratives in each category were then analyzed for levels of descriptive detail in history of illness and physical exam (HPI PE), diagnostic findings and detail of ED course.

Results
About a half of handoff narratives in each category contained only minimal descriptive detail: 53% of HPI, 47% of diagnostic findings and 55% of ED course information. Detailed descriptions of HPI were communicated for 16% of patients, 3% of diagnostic findings and 12% of ED course information – the smallest proportions in each category. There were no descriptions of diagnostic findings for 31% of cases and 16% of ED course. Diagnostic uncertainty was classified into three levels: 1 - Clear presumptive diagnosis (most confident), 2 - Narrow differential and 3 - Unclear (least confident). The most frequently observed level was 2 (51%). For 13% of patients the diagnosis was Unclear. Fit to script was determined as: 1 – typical (41%), 2 – Variant of a typical problem (19%) and 3 – No recognized fit (9%). Comparing content across the categories of uncertainty we found that levels of descriptive detail were largely determined by confidence in diagnostic judgment. As uncertainty increased, physicians provided more information in handoff presentations. Statistical analyses showed significant difference between groups classified by uncertainty.

Conclusion
When patients do fit well onto an established process script, physicians do not see value of adding additional detail to the handoff. Effective electronic support tools should allow for variations in complexity and diagnostic certainty.

Stakeholder Perspectives on Policy Implications
Post the Conclusion of the State HIE Program

Lauren Hovey, MA1, Petry Ubri1, Prashila Dullabh, MD1
1NORC at the University of Chicago, Bethesda, MD

Research Objective
Since the passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act, and with it the creation of the State HIE Program and other health information exchange (HIE)-supportive initiatives, adoption of electronic health records (EHRs) has grown significantly and HIE has expanded. Following the conclusion of the State HIE Program, there are a number of opportunities at the state and federal levels to continue enabling HIE services.

Study Design
We draw these policy implications from three evaluation activities: a qualitative, in-depth assessment of eleven states consisting of site visits; semi-structured stakeholder discussions in five states, and 60-minute summative interviews with key informants with diverse perspectives on HIE. From these activities, we gleaned important lessons learned and suggestions for maintaining and expanding HIE following the conclusion of the program. For the case studies, NORC held discussions with more than 100 individuals in eleven states. For stakeholder discussions, NORC had conversations with forty-five stakeholders in five additional states. Informants included providers, payers, vendors, Medicaid and Public Health personnel, and state program leadership, among others. The summative key informant discussions included 22 thought leaders across state, federal, and private sectors.

Principal Findings

- States play an important role in leadership and coordination, particularly in convening stakeholders, policy development, and needs assessment, which will be relevant to future stages of meaningful use and health care reform.
- There is a critical need for strong, ongoing federal support related to governance, technical standards, and interoperability.
- There is a need for alignment of HIE efforts across federal agencies to leverage HIE investments. The healthcare system’s HIE goals require the development and use of consistent technical standards for information flow which could potentially be pursued by an aligned “policy push” with federal incentives for standards of information.
- A provider- or federal-led effort to obtain buy-in from HIE developers for overarching HIE goals may be warranted and needed to change perceptions of interoperability.
- There is a need to assess how technical solutions evolve in different markets, to develop and disseminate best practices, and to develop governance and oversight requirements at a state and national level. This is a responsibility best shared between state and federal leaders.
- Organizations should monitor findings and best practices for sustainability from existing federal and state level initiatives leveraging pay-for-performance models.

Conclusions
The results suggest that federal- and state-level entities have important roles to play in the next phase of HIE. Continuing the expansion and maximizing the utility of electronic data exchange will require their leadership and guidance in terms of policies, data standards, and best practices. Robust and sustainable HIE also requires participation, buy-in, and needs assessments from healthcare organizations, providers, and developers.

Implications for Policy, Delivery or Practice
HIE sustainability will be a collective effort with important roles for stakeholders and added support from a new generation of initiatives, such as the Health Care Innovation Awards and State Innovation Model funding.
Evaluation of a Local Terminology to SNOMED CT Crosswalk

Rachael Howe, MSN, RN, Tiffany Harman, RN, Pat Wilson, RT(R), CPC, PMP, Lee Min Lau, MD, PhD 3M Health Information Systems, Inc., Murray, UT

Introduction: The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a clinical terminology used to represent clinical concepts in healthcare for the purpose of increasing interoperability between electronic systems. The goal of this study was to evaluate the crosswalk created between two terminologies, the target being SNOMED CT and the source being a local terminology, to see if the mapping was accurate, so that data encoded with the local terminology could be translated to SNOMED CT for data exchange. The local terminology contained multiple clinical concepts across the full spectrum of healthcare, effectively resulting in the need to utilize all of SNOMED CT for comparison.

Method: There were 3 phases in this project: analysis of the local terminology content, examination of the relationships between each source and target concept, and evaluation of findings. Analysis consisted of studying the documentation provided along with the data to understand how and why the data was mapped. The local terminology content was then assessed to understand how the relationships were determined. Next, each concept to concept association was examined in depth. Finally, the overall data was reviewed for inconsistencies and gaps in concept coverage and meaning, and to establish metrics.

Results: The crosswalk utilized 62 relationship types, of which 59 mapped to attributes (which are in turn members of the “linkage concept” high level domain) in SNOMED CT. However, two of the relationship types used in the crosswalk mapped to the same SNOMED CT concept ID, thus only 58 SNOMED CT attributes were used in the crosswalk. In addition, 3 new relationship types (not found in SNOMED CT) were created for the crosswalk. The documentation provided with the data did not indicate the reason for each relationship type nor how they were to be used. Thus each concept-relationship-concept triplet was evaluated solely on the definition implied by the description or designation, the majority of which were understandable. There were a few instances, however where semantic drift may have occurred. For example, the SNOMED CT attribute of ‘Is A’ (SNOMED CT ID 116680003) was altered to “Narrower than (IS A)”, which may have a negative impact on the accuracy of the local concept mapping to SNOMED CT. This is especially significant since this relationship type accounts for the largest proportion of mapping rows (over 39%, or 166,125 out of 424,488 rows). Another example is the inverse relationship of ‘Broader than’, which was not derived from an existing SNOMED CT attribute, is a misnomer for a grouping of local terms that have a single property or concept component in common (see Table 1 below).

<table>
<thead>
<tr>
<th>LOCAL CONCEPT</th>
<th>RELATIONSHIP TYPE</th>
<th>SNOMED CT PREFERRED TERM</th>
<th>SNOMED CT ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retina</td>
<td>Broader than</td>
<td>O/E - no retinopathy</td>
<td>390847009</td>
</tr>
</tbody>
</table>

Table 1: Example of a ‘Broader Than’ Relationship

Conclusion: Our final analysis showed that although a lot of local terms could be crosswalked to SNOMED CT through the mapping created, there were cases where the relationship types applied were not consistent in meaning and usage. Existing informatics principles should be considered and documented when establishing a new crosswalk or map to ensure that it is meaningful, useful and reproducible. This particular crosswalk would require additional review and revision to improve accuracy in mapping from the local terminology to SNOMED CT, if the SNOMED CT codes were to be used instead of the local codes in external data exchange, to ensure that the meaning of the patient data remains unchanged.

Migration of a Computerized Anticoagulation Clinic to a Commercially-Developed EHR

Ping Hu, MD, MS1, Mitchell Hedrick2, Aubrey Ramsey, MS2, Christopher Radek2, Keith Huffman2, Adam Miklius2, Adam Wilcox, PhD1
1Intermountain Healthcare, Salt Lake City, Utah; 2Cerner World Headquarters, Kansas City, MO

Abstract: We describe the migration of a computerized anticoagulation clinic application from an internally-developed clinical information system to a commercially-developed EHR. Specific functions are noted, with the corresponding EHR tools that supported them. Initial statistics show successful implementation of the system, which was developed using only existing tools in the commercial EHR.

Introduction: Intermountain Healthcare, long known for its use of clinical decision support (CDS) in an internally-developed electronic health record (EHR), is migrating to Cerner Millennium. An important concern is how advanced CDS can be successfully migrated to a commercially-developed EHR.

Background: In their 2006 landmark review of the impact of health information technology (health IT), Chaudhry et al. identified 4 benchmark institutions, all with internally-developed EHRs, that contributed much of the research on health IT impact. At the same time, they noted a dearth of research on the impact of commercially-developed EHRs. Intermountain recently purchased Cerner EHR, and there is a critical need to migrate these successful CDS applications from an internal, flexible development environment to a more-constrained, commercially-developed system. Recent studies have shown clinicians’ dissatisfaction with commercial EHRs, perhaps due to decreased flexibility. At the same time, internally-developed EHRs are becoming extinct. We describe our migration of a computerized management of an anticoagulation clinic application (CAC) to the Cerner Millennium EHR. This represents an important instance of migrating a successful and complex CDS application to a commercial EHR.

Methods: The CAC application had the following functions: protocol enrollment; INR result follow-up alerting and documentation; warfarin dosage calculation; drug interaction checking; alert routing and acknowledgment; and INR result graphing. The Cerner Millennium components used for each function were: Order Entry (protocol enrollment, drug interaction checking), PowerChart (protocol enrollment, INR result documentation, alert acknowledgment), DiscernDev (INR result follow-up alerting, drug interaction checking, alert routing and acknowledgment), and mPages (warfarin dosage calculation, INR result follow-up and documentation, INR result graphing). A multidisciplinary team, including clinical users and developers, participated in the analysis, design, development and testing of the system in Cerner Millennium.

Results: The application was developed over a 6-month period. There were 217 rules developed, and 11 algorithms. The application was first released on 2/21/2015. Six clinics enrolled 746 patients in the protocol in the first 19 days – these included many patients who had been previously enrolled in the legacy system. There were 1397 alerts generated over the same period. Most of the functions were supported by the standard commercial tools, including the specific data-driven and time-driven rule triggering, alert routing, and alert escalation. Some functions needed more advanced tools within the commercial EHR development environment, but these were supported by the newer mPage development tools. The flexibility of the mPages was critical for the more complex components of the application. No development was needed outside the standard tools or mPages.

Discussion: We were able to successfully migrate a complex CDS application from our internally-developed EHR to a commercially-developed EHR. The original CAC application had been developed and configured over a 10-year period, as guidelines were adjusted and workflows adapted, so it was critical that this optimized functionality was supported. The application is used in over 40 Intermountain clinics, for over 5,000 patients. The existing Cerner development applications were sufficient to support the development, but only with the inclusion of mPages. We expect as mPage functionality advances, and additional sharable health services are developed, even greater flexibility will be supported in Cerner Millennium for complicated CDS.

References
Deep Sequencing of Phage-Displayed Peptide Libraries Reveals Novel Peptide Motif that Detects Norovirus

Amy M. Hurwitz, B.S., Wanzhi Huang, Ph.D., Robert L. Atmar, M.D.,
Timothy G. Palzkill, Ph.D.
Baylor College of Medicine, Houston, TX

Introduction

Norovirus (NoV) infections are the most common cause of non-bacterial gastroenteritis and lead to about 21 million new cases and cost $2 billion per year in the United States alone. Existing diagnostics have limited feasibility for point-of-care applications, so there is a clear need for more reliable, rapid, & simple-to-use diagnostic tools for detecting NoV. Previous studies have provided proof-of-concept for the use of bacteriophages, or bacteria-infecting viruses, to detect noroviruses. Here, we use deep sequencing and computational analysis to identify a consensus motif within a library of phage-displayed peptides that bind norovirus in vitro.

Methods

In this study, phage display technology was used to screen a naïve M13 phage library displaying over a billion randomized 12-mer peptides for those that bind to NoV virus-like particles, or VLPs (Figure 1). Since NoV strains classified in genogroups GI and GII cause the majority of human outbreaks, we focused on the prototypical Norwalk strain (genotype GI.1) to identify binding reagents. Phage populations were characterized by deep sequencing after three and four rounds of selection for NoV binding, and custom scripts were used to perform computational analysis for the identification of amino acid motifs involved in binding GI.1 VLPs.

![Phage Display Flowchart]

Figure 1. One round of selection involves incubation, washing, elution, and amplification for subsequent rounds.

Results

Deep sequencing of selected phage populations revealed thousands of unique peptides with possible binding to GI.1. The peptides occurring most commonly increased in fractional frequency from round three to four, as did the overall binding affinity of each population for GI.1 VLPs. Computational sequence analysis identified consensus motif, “YRSWXP,” from the phages obtained after round four of selection. Phage ELISAs and solid-phase peptide arrays confirmed the ability of this motif to bind specifically to GI.1 in the context of two 12-mer peptides.

Conclusions

The work reported here provides evidence for the use of phage display technology and deep sequencing for the identification of novel peptides that detect NoV. This method enables rapid identification of lead clones that bind new pathogens and can be further developed to produce optimized protein binders. Further, using phage-displayed peptides directly as detection reagents provides advantages including the ease and low cost of phage propagation and purification, and high avidity provided by the phage structure. Ultimately, this work will lead to the establishment of novel tools for rapid and sensitive diagnosis of NoV infection with specificity for a broad range of clinically relevant NoV strains.

References

ArticlesAboutMe.org: Disseminating Clinical Trials Results to Patients

Vojtech Huser, MD, PhD1, Anil Yaman MS2, Chunhua Weng PhD2, James J Cimino MD3

1Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD; 2 Columbia University, New York, NY; 3University of Alabama-Birmingham, Birmingham, AL

Introduction
Advances in Web technologies present new opportunities to keep clinical trial participants informed about trial results. Increased interest of participants and trial sponsors in results transparency is reflected in the “Principles for Responsible Data Sharing” issued by the Pharmaceutical Research and Manufacturers of America (PhRMA) in July 2013.1 Prior reports indicate that many human clinical studies terminate due to lack of enrollment.2 Promising to keep participant informed may increase their participation in clinical trials.

Methods
To encourage and reward clinical trial participation by patients, we created a Web service (available at www.ArticlesAboutMe.org) that generates an email every time an article related to a clinical trial is published. The service relies the ClinicalTrials.gov registry (using the NCT identifier associated with the trial) and the PubMed article repository. A subscription to trial email updates is initiated by the trial participant or the trial administrator (after preauthorization). The service performs weekly checks for result publication using two structured trial-article links3 and a free text search for trial NCT.

Results and Discussion
The service was created in October 2014 and piloted until December 2014 using Amazon Web Services cloud infrastructure and the .NET web technology. Since official launch in January 2015, the service has been used for monitoring of over twelve trials. A total of 9 emails with clinical trial results were sent during 11 completed weekly scans (as of March 2015). The service was designed for not only actual trial participants but also general health care consumers that are interested in receiving results of a trial of interest. The majority of clinical trial result publications were detected using free-text PubMed NCT search.

A marketing campaign to selected principal investigators of clinical studies revealed only limited interest to use the service. Targeting trial participants directly at an institutional level or engaging patient advocacy groups may prove more effective. The poster will show results of pending service integration at Columbia University’s Irving Institute for Clinical and Translational Research.

References
Examining the Role of Bug-tracking Systems in the Maintenance of Electronic Health Records (EHRs)

Salman T. Hussain1, Dustin S. McEvoy2, Thu-Trang T. Hickman, MPH1, Dean F. Sittig, PhD3, Adam Wright, PhD1,2

1Brigham and Women’s Hospital, Boston, MA; 2Partners HealthCare, Wellesley, MA; 3University of Texas at Houston Health Science Center, Houston, TX

Introduction
The maintenance of electronic health records (EHRs) requires the intersection of the healthcare and software engineering industries. Many teams are working on the same project, as software developers, analysts, and physician end-users are all seeking to improve the quality of the EHR. Communication and coordination problems between these groups must be avoided1,2, as they may put patients at risk. For this reason, engineers at Partners Healthcare designed the Team Coordination System (TCS), a computer program with an integrated software team environment enabling different groups to report and track software bugs, application enhancements, changes in specifications, code control, versioning, and releases in the EHR through its dialogue function.

Methods
We conducted a literature review to assess bug-tracking systems used in the monitoring of EHR systems. An initial query of all closed and fixed TCS issues in the Partners’ Longitudinal Medical Record yielded 12,996 issues from 4/10/02 – 3/2/15. We calculated the mean and median resolution times of issues, and also the proportion of issues by their type. We separately reviewed over 1500 TCS issues by querying closed and fixed issues containing CDS-related terms such as “alert” and “reminder”, specifically focusing on how bugs were monitored, to understand the ability of TCS software to help report, identify, and track bug fixes in the EHR.

Results
We observed that the mean resolution time for EHR issues was inversely correlated to their severity. The mean resolution time was lowest for the most severe level 1 issues (11.0 days; median 3 days), though we were surprised that even some of these highest severity issues required up to 144 days to fix. The mean resolution time for level 2 issues was longer (55.3 days; median 10 days) and even longer for level 3 issues (66.3 days; median 23 days). Low severity level 4 issues (89.7 days; median 45 days) and the least severe level 5 operational issues (109.1 days; median 57 days) had the longest resolution times. We also analyzed issues by report type and found that 63.5% of EHR issues in TCS were bugs, 15.0% support issues, 14.4% enhancements, 4.1% usability issues, 1.6% specifications, 0.8% documentation requests, and 0.6% report requests.

Conclusion
While bug-tracking systems are likely used in the design, development, and testing of most EHRs, to our knowledge there are no publications analyzing their use in this context. Retroactively mining bug reports can help us learn how to make systems safer. End users do not currently have access to TCS, as they simply call the help desk and open a ticket to report an issue. Granting end users read/write access to TCS could reduce dependencies between groups, potentially leading to safer and more effective EHRs, not to mention more rapid identification and solution of problems. This analysis highlights that TCS is an instrumental tool to track and fix bugs in EHRs, but that it could be improved with increased usability for physicians.

References
2. Begel A, Nagappan N, Poile C, Layman L. Coordination in Large-Scale Software Teams. In: Proceedings ECSE Workshop on Cooperative and Human Aspects on Software Engineering (CHASE); May 2009; Vancouver, BC.
Examining Cancer Case Reporting Processes and Timeliness: Preliminary Results

Abdulrahman Jabour¹, Brian E. Dixon²,³,⁴, Josette F. Jones¹, David. Haggstrom²,⁴,⁵

¹ Department of BioHealth Informatics, Indiana University School of Informatics and Computing; ² Regenstrief Institute; ³ Department of Epidemiology, Indiana University Fairbanks School of Public Health; ⁴ Indiana University, School of Medicine; ⁵ HSR&D Service, Veterans Health Administration, Department Veterans Affairs

Introduction

Cancer registries collect tumor-related data to identify incidence rates, monitor trends, and support cancer research. One of most common issues with cancer data is timeliness. Delay in reporting is an ongoing, structural issue that is discussed in several U.S. Centers for Disease Control and Prevention (CDC) and Institute of Medicine (IOM) reports. Very few published studies within the United States have examined the reporting timeline within cancer registries.

Often the data collected in local hospitals and other treatment facilities are stored in the hospital’s local registry which are subsequently reported to the state registries. Subsequently, state registries may share data with larger national registries such as Surveillance, Epidemiology, and End Results Program (SEER), North American Association of Central Cancer Registries (NAACCR) and CDC's National Program of Cancer Registries (CDC/NPCR). Delays or bottlenecks in reporting processes at the hospital or state registry level could lead to delays in reporting at the national level. Meeting national standards will have both financial and credibility implications. Therefore, to meet the need for actionable data, both reporting hospitals and the state registry should consider way to accelerate reporting timeliness. Studies in some European countries such as Norway and Sweden show an improvement in timeliness over the years. In this study, we evaluate the timeliness of data at the Indiana State Department of Health’s (ISDH) Cancer Registry from 2001 to 2009. The study will also examine how ISDH Cancer Registry data compares to other, national registries’ standards.

Method

Reporting timeliness was calculated by measuring the number of days from the date of diagnosis to the date when data become available for reporting at the state registry. The study includes breast, colorectal, and lung cancer cases in the state of Indiana from 2001 to 2009. Calculations include the mean, median, and the percentage of records that meet the predefined measures for reporting performance. Predefined measures was drawn from the following national standards for cancer registries; SEER, NAACCR, and, CDC/NPCR. Delays in reporting were stratified by the type of cancer and by year to assess the degree of consistency in reporting speed.

Preliminary Result

A total of 66,409 cases were examined. Depending on year of diagnosis only 3.9% to 37.2% of cases were reported within 6 months from the date of diagnosis. Between 94.8% and 98.8% of cases were reported within 2 years. The CDC/NPCR requires 95% or more of the cases to be reported within 24 months from the date of diagnosis, and SEER requires 98% or more to be reported within 22 months. NAACCR requires 90% or more for silver, and 95% or more for gold within 23 months. Comparing these standards, the ISDH Cancer Registry met the CDC/NPCR and NAACCR for the entire 9 years, but it only meet the SEER standard in 2002, although the ISDH tumor registry does not participate in the SEER program. The result also shows that reporting timeframes did not vary meaningfully based on cancer type.

Over time, we observed a large variation in the mean reporting time across time that ranged from 426 days in 2003 to 252 days in 2009. Following an initial sharp (46%) increase in reporting time from 2002 to 2003, we observed a gradual decline in reporting times through 2009. The longer reporting time observed in 2003 was concurrent with changes in the reporting requirements such as the implementation of the Facility Oncology Registry Data Standards (FORDS) coding standard.
Representation of Clinical Practice Guideline Data Elements Using the Health Level Seven Fast Healthcare Interoperability Resources (FHIR) Standard as a Proposed Data Formalism for the Arden Syntax

Robert A. Jenders, MD, MS, FACP, FACMI
Charles Drew University & University of California, Los Angeles, CA

Abstract
Context: Arden Syntax can adequately represent clinical practice guidelines (CPGs) in clinical decision support systems but lacks a standard data formalism. Objective: Assess Health Level Seven’s Fast Healthcare Interoperability Resources standard (FHIR) representation of CPG data. Method: 16 CPGs containing 806 data elements were tabulated. Result: FHIR can represent all but 2.7% of these data elements. Conclusion: FHIR adequately represents data elements in CPGs. Incorporation of FHIR in Arden Syntax would facilitate representation of CPGs.

Introduction
Arden Syntax is an American National Standards Institute (ANSI) standard supervised by Health Level Seven (HL7) for representation of executable medical knowledge with the goal of facilitating its sharing among clinical decision support systems (CDSS). Key to minimizing site-specific changes when sharing knowledge is the standardization of database linkages, which in turn requires a standard data model, vocabularies and query syntax that are not present in Arden1. The HL7 FHIR draft standard for trial use (DSTU) Release 1 is a new framework that allows references to data to be represented in XML and bound to terminologies in modular components known as “resources.” For example, concepts such as “patient,” “medication” and “observation” are key FHIR resources, each with structured attributes that may be other resources. The American College of Physicians (ACP) is a large professional organization of internal medicine specialists in the USA that publishes evidence-based clinical practice guidelines (CPGs) pertinent to medical practice that are representative of CPGs generally because of ACP’s size and clinical focus. The present work was undertaken to assess FHIR’s utility in representing data elements in ACP CPGs that in turn might be represented in a CDSS using Arden Syntax in order to assess FHIR’s utility as a standard data formalism for the Arden Syntax.

Methods
The data elements in the current ACP CPG library were identified and tabulated using Microsoft Excel. Each then was assessed to ascertain whether it could be specifically represented using FHIR DSTU Release 1.

Results
Three hundred sixty-eight distinct data elements were identified among 806 total elements in the CPGs that contained 121 recommendation statements that could be implemented in a CDSS. Nearly all of these elements could be classified as observations, procedures, medications, diagnoses and diagnostic reports and could be represented in FHIR. Of the 368 distinct data elements, 11 (3%) corresponding to 22 (2.7%) elements in the collection of 806 total elements could not be directly represented. These included elements such as “preference” and inference data such as “ease of use.”

Conclusions
FHIR is adequate to represent the data elements found in a robust corpus of CPGs that in turn would be a typical knowledge base in a CDSS. Consideration should be given for use of FHIR to represent data elements in a standard way in the Arden Syntax in order to facilitate knowledge sharing.

Acknowledgements
This study was supported in part by the US NIH, grants UL1TR000124 (NCATS) and 2U54MD007598 (NIMHD).

References
An Interactive System for Comprehensive Geriatric Telerehabilitation

In cheol Jeong, PhD¹, Joseph Finkelstein, MD, PhD¹
¹Chronic Disease Informatics Program, Johns Hopkins University, Baltimore, MD

Introduction

High acceptance of home-based telecare applications by patients with various chronic health conditions has been widely demonstrated. In older patients, particular interest has been expressed in home-based telerehabilitation supporting regular exercise. To address this interest, we developed an interactive aerobic and strength exercise system based on Home Automated Telemanagement (HAT) model. In this article, we present the system design and introduce the system performance.

System Design

The design of the home-based telerehabilitation system supports flexible exercise prescription and remote control of exercise settings by a comprehensive rehabilitation team, automated exercise records, real-time monitoring of patient progress towards recommended patient-specific exercise goals, and generation of exercise safety alarm. The home station, such as PC, tablet or smart phone, generates a detailed exercise log which can be communicated to the patient rehabilitation team. This allows the patient providers determine which exercise settings are more challenging to the patient. There are also general exercise safety tips as well as specific tips for each exercise designed to minimize the risk of injury during exercise. A touch screen dashboard is used for the patient interface to start, stop or skip exercise by the patient preferences and to receive feedback on the exercise status during patient exercise. The dashboard has been optimized for individuals with possible limitations in vision, locomotion, and cognition. The patient unit includes an arm/leg biking device (SF-B02 Motorized Mini Bike, Sunny Health Fitness, USA) and a resistance chair with attached elastic bands (CFC-100B-S, VQ ActionCare, LLC, USA) and SpO2 monitor that sends heart rate (HR) and SpO2 level data to the home station. The system design is shown in Figure 1.

System Performance

The system can acquire all exercise and physiological information throughout execution of prescribed exercises. During the aerobic exercise, current and target information such as speed, distance, exercise time, calories burnt, HR and SpO2 level are presented in a user-friendly trending format on the exercise dashboard in a real-time mode. During the resistance exercise, type, direction, target, number of repetitions, percentage of exercise similarity, HR SpO2 levels and whether the user reaches the target goals are presented. The entire log of the performed exercise is directly sent to the HAT server. When any exercise parameter exceeds the pre-setup alarm thresholds, the patient is notified and the program ends automatically for the entire exercise session. Additionally, the aerobic exercise can provide alternative exercise modes including patient-driven mode for fit patients and assisted motor-driven mode for frail patients.

Conclusion

Comprehensive home-based geriatric telerehabilitation system has significant potential in providing aerobic and strengthening exercise support and performance feedback, communicating results to a clinical rehabilitation team, and supporting safe and efficient exercise at senior patient homes.

Figure 1. Home-based telerehabilitation system
Harmonization of Quality Data Model with HL7 FHIR to Support EHR-driven Phenotype Authoring and Execution: A Pilot Study

Guoqian Jiang¹, Harold R. Solbrig¹, Richard Kiefer¹, Luke V. Rasmussen², Huan Mo³, Jennifer A. Pacheco², Enid Montague², Jie Xu², Peter Speltz³, William K. Thompson², Joshua C. Denny³, Christopher G. Chute⁴, Jyotishman Pathak¹
¹ Mayo Clinic College of Medicine, Rochester, MN; ² Northwestern University, Chicago, IL; ³ Vanderbilt University, Nashville, TN; ⁴ Johns Hopkins University, Baltimore, MD

Abstract
The objective of this pilot study is to describe a crowdsourcing effort in harmonizing high-level data elements between Quality Data Model (QDM) and HL7 Fast Healthcare Interoperability Resources (FHIR) to support electronic health records (EHR)-driven phenotype authoring and execution. In total, 206 mapping pairs between the two models were identified, with a Fleiss's kappa statistics (k=0.24) calculated for inter-rater agreement. We discuss challenging issues of the mappings.

Introduction
In previous studies, we developed a metadata repository and its services for data elements derived from Quality Data Model (QDM) and HL7 Fast Healthcare Interoperability Resources (FHIR) to support electronic health records (EHR)-driven phenotype authoring and execution (1-2). In the present study, we describe a crowdsourcing effort in harmonization of high-level data elements between the two models. The purpose of the harmonization is to understand the domain coverage of the two models and to establish the degree of interoperability between QDM-based phenotype algorithms and patient data populated with FHIR models.

Methods
We extracted a collection of high-level data elements with their textual definitions from both QDM and HL7 FHIR using our metadata repository services (http://projectphema.org). We designed a mapping application using an Excel spreadsheet. We created five worksheets: instructions, mappings, QDM data elements with definitions, FHIR data elements with definitions, and SKOS mapping properties with definitions. One-to-many mappings are allowed by the spreadsheet application. We used five SKOS mapping properties (exactMatch, closeMatch, broadMatch, narrowMatch and relatedMatch) as the mapping types. We asked the project team members to complete the mappings individually and Fleiss' kappa statistics was calculated to assess inter-rater agreement.

Results and Discussions
In total, 94 data elements from QDM (consisting of 18 QDM Categories and 76 QDM Datatypes) and 98 data elements from FHIR (all FHIR Resources) were extracted for mappings. We received the responses from 7 team members and 206 mapping pairs were created. All QDM data elements had at least one mapping suggested whereas 65 FHIR data elements did not have any mappings. Only fair agreement (kappa = 0.24) was achieved. The QDM categories Communication, Condition/Diagnosis/Problem and Encounter were in relatively high agreement. The textual definitions of the data elements, along with multiple factors (e.g., attributes associated with each type, ambiguity in naming within two models), are important for creating correct mappings. In summary, the pilot study provides valuable insight into the challenging issues for a community-based harmonization of QDM and HL7 FHIR. In the future, we plan to extend the harmonization effort in collaboration with broader clinical research communities, aiming at producing a collection of complete and high-quality mappings.

Acknowledgments: This work has been supported in part by funding from PhEMA (R01 GM105688), eMERGE (U01 HG006379, U01 HG006378 and U01 HG006388), and caCDE-QA (1U01CA180940-01A1).

References

1512
Physician Participation in Meaningful Use and Rehospitalization of the Dually-Eligible

Hye-Young Jung, PhD1, Mark Aaron Unruh, PhD1, Joshua R. Vest, PhD1, Lawrence P. Casalino, MD, PhD1, and Rainu Kaushal, MD, MPH1,2,3,4, for the HITEC Investigators


Abstract: Dual-eligibles have higher rates of rehospitalization compared to the broader Medicare population. Outpatient physicians’ use of interoperable electronic health records may reduce readmission rates. We compared the odds of rehospitalization pre/post Meaningful Use for dual-eligibles attributed to physicians who participated in the program with dual-eligibles attributed to matched control physicians. Relative to the control group, physician participation in Meaningful Use was associated with 6% lower odds of rehospitalization (Odds Ratio: 0.94, 95% confidence interval: 0.90-0.98).

Introduction:
Compared to the broader Medicare population, dual-eligibles participating in both Medicare and Medicaid have substantially higher rates of rehospitalization.[1] Readmission has been associated with gaps in care following hospital discharge,[2] a common problem for dual-eligibles. Physician use of interoperable electronic health records (EHRs) may reduce rates of rehospitalization of dual-eligibles. Use of highly functional EHRs within a practice can help with outpatient coordination, and interoperable exchange of information with hospitals can help with transitions of care to and from the hospital. The Meaningful Use program has made nearly $30 billion available to encourage provider adoption and use of interoperable EHRs.[3] The objective of this study was to evaluate the impact of outpatient physicians’ participation in Meaningful Use on rehospitalization of dual-eligibles.

Methods:
Multiple sources were used to construct our analytic file. Medicare enrollment records, carrier files and MedPAR files were used to derive patient demographics, diagnosis and procedure codes, hospital length of stay, and use of an intensive care unit. Physician and practice characteristics were obtained from the SK&A Physician Database. Meaningful Use participation was identified through dates of successful attestation found in program payment records. The study population included 316,468 dual-eligibles in the State of New York hospitalized over the period 2010-2012. Each patient was attributed to a physician with a plurality of outpatient evaluation and management claims. We used a quasi-experimental design, utilizing the differential timing of physician participation in Meaningful Use to estimate the program’s effect on rehospitalization. Our outcome variable was a binary measure of all-cause 30-day readmission. The determinant of interest was a binary indicator of Meaningful Use participation. The intervention group included patients attributed to physicians who participated in Meaningful Use during the study period. Comparison groups included patients attributed to physicians who joined Meaningful Use after the study period, but who used paper records or EHRs without Meaningful Use participation during the study period. Estimates were derived with a multivariable logistic regression model with hospital and year fixed effects that controlled for temporal trends and unobserved cross-sectional heterogeneity between hospitals.

Results:
Physician participation in Meaningful Use was associated with 6% lower odds (Odds Ratio: 0.94, 95% confidence interval: 0.90 to 0.98) of rehospitalization among attributed dual-eligibles compared to patients attributed to physicians using paper records or EHRs without Meaningful Use participation.

Discussion:
Our findings suggest that investments in interoperable EHRs made through Meaningful Use may improve care for dual-eligibles through reductions in rehospitalization. Our study represents the first evidence of the impact of physician participation in the program on the quality and efficiency of care among the dually-eligible. Future study is warranted to examine whether further reductions in readmission rates will occur as physicians gain more experience with EHRs and as interoperable exchange of information with hospitals increases. It will also be important to determine whether a similar reduction in readmissions occurs for patients of physicians who join the Meaningful Use program in later years.

References
Natural Language Processing facilitates delivery of individualized recommendations at the point of care

Mayo Clinic, Rochester, MN, USA

Introduction
Mayo Clinic has made significant strides in developing knowledge assets and the tools to manage them. To fully realize the value of these assets, it is pertinent not only to embed the knowledge into the clinical workflow but also contextualize the knowledge to individual patients\(^1\). To ensure the delivery of standard guidelines, we have created care process models (CPMs) for various conditions. However, individualized patient care according to CPMs requires the review and obtainment of relevant clinical information from both structured data fields and unstructured sources of information\(^2\) such as clinical notes, lab reports, etc. While information in structured fields is highly useful, the integrated information from both structured and unstructured repositories at various time points provides comprehensive views during personalized decision-making. Here, we discuss a multi-stage iterative development of an NLP system that extracts data elements needed for automating care process models for hyperlipidemia, atrial fibrillation and heart failure. We also investigated how NLP helps to fill the information gap that exists in structured data sources and empowers clinical decision-making at the point of care.

Methods
As the first step, we developed a gold standard corpus based on the guidelines (Supplementary File1\(^3\)). The corpus was identified through stratified sampling from a retrospective cohort for the data elements pertaining to the three care process models (CPMs) namely Atrial Fibrillation, Heart Failure and Hyperlipidemia. Based on the gold-standard annotations from clinical notes, we developed regular expressions using MedTagger\(^4\) to extract (See Supplementary File2\(^5\)) the data elements from selected sections in clinical notes. The choice of a rule-based approach stems from the overall goal of the project to achieve a very high precision (~100%), at the cost of lower recall. We iteratively refined the rules developed to extract data elements based on the analysis carried out after each run.

Results
The system was evaluated against the gold standard annotated data set in an iterative manner. The system achieved an overall F-Measure of 92.61% on blind data set. We also conducted a pilot evaluation regarding the impact of NLP on individualized care recommendations made on 14,000 patients. The NLP system decisions when combined with the information in structured data made significant differences to the care recommendations to 1,730 patients\(^6\). The NLP system reversed the care recommendations for 155 patients, while it suggested alternative care recommendations to 1, 885 patients. We will soon realize the robustness of the NLP recommendations as the physicians are reviewing the care recommendations made by MEA in the pilot study.

Conclusions and Future work
We have successfully integrated an NLP system into an EMR agnostic care recommendation solution system, MayoExpertAdvisor. The adopted batch-wise iterative system development has yielded a robust NLP system with acceptable performance.

References
Selection of a Database Engine for the National Master Patient Index of Malawi

Chimwemwe Kachaje1, Blessings Mhango1, Mwatha Bwanali1, Soyapi Mumba2, Jeremy U. Espino3 MS MD, Gerald P. Douglas3 PhD
1Baobab Health Trust, Lilongwe, Malawi, Africa
2Center for Health Informatics for the Underserved, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Abstract

Objective: Select a database engine for the Malawi national master patient index that supports multiple synchronized databases over an unstable network. Methods: We reviewed the capabilities of the database engines that fulfill the objective using multi-master replication. We tested the selected database using a 22 test cases and simulated dataset of 72.9 million records. Results: We chose CouchDB as our database solution and it passed our test cases. Conclusions: CouchDB can provide an infrastructure for a distributed, scalable and reliable master patient index.

Introduction

On March 19th 2001, we deployed the first Baobab Health Electronic Medical Record (EMR) system at Kamuzu Central Hospital in Lilongwe, Malawi.1 In 2009, the EMR was adopted nationally for the HIV clinics at district and central hospitals of Malawi. As usage grew, users of the system wanted consistent and seamless access to patient demographics across the district and central hospitals.2 To fulfill this need we needed a master patient index (MPI).

Initially, we rolled out an MPI solution based on a centralized MySQL database and a data synchronization solution that we wrote ourselves, the Demographic Data Exchange (DDE). As the number of sites using the DDE grew to 29 sites, it became apparent the system could not reliably synchronize data. The DDE suffered from denial of service when multiple clients tried to synchronize simultaneously. When there were extended network outages, data synchronization among the sites did not finish and could not recover. The DDE issued duplicate IDs because it did not handle concurrent operations. Over time, data from one site was inconsistent with the data at another site.

We recognized that the challenges we encountered with the first version of the DDE were because of the complexity of implementing a multi-master replicated database cluster. Multi-master replication is a method of database replication which allows data to be stored by a group of computers, and updated by any member of the group.3

Methods

We examined the characteristics of existing open source databases that support multi-master replication—CouchDB, MySQL, and PostgreSQL. We wrote and executed a test plan for testing the synchronization capabilities of the selected database.

Results

We selected CouchDB as our database solution. Our research revealed that although MySQL supports multi-master replication it requires a stable communications network to operate well.4 Multi-master replication for PostgreSQL would not appear until PostgreSQL version 9.4.3

We wrote a set of 22 test cases and generated a simulated data set of 72.9 million records (the maximum number of patients we could represent using our seven character base 30 patient id). We executed the test cases in our laboratory using the data set—evaluating consistency of the data over simulated network conditions. An example test case was “Test #16 – Synchronization of IDs and Benchmarking of Cellular Connection; Procedure: Allocate 10 IDs on node 1. Perform synchronization to nodes 2-5. Record synchronization time; Test Criteria: Synchronization time recorded. 10 NPIDs are present/updated on all proxies and master.”

Conclusion

CouchDB has the potential to provide an ideal infrastructure for a national master patient index. Such a system would permit patients to use a single ID for all facilities, enhance monitoring and evaluation statistics and in the future, share abstracts of medical records across sites.
Implementing an EMR-based “No Opiate” Policy In The Emergency Department

Zachary P. Kahler, MD¹, Paul I. Musey, MD¹, John T. Finnell, MD MS², Annelyssa Johnson, MD¹, Christi Saupp, RN¹, Christian Strachan, MD¹, Charles Shufflebarger, MD¹
¹Indiana University, Indianapolis, IN, USA; ²Regenstrief Institute, Indianapolis, IN, USA

Introduction
Opiate-seeking behavior is a major problem in U.S. emergency departments (ED) and leads to recidivism and ED misuse¹,². Our department has created an EMR-based opiate “warning flag” to help transition frequent opiate-seeking patients away from the ED and into an outpatient chronic pain program.

Outcomes
The primary objective is to evaluate the effect of the EMR-based intervention on patient visits to the ED. Secondary outcomes include number of lab draws, ECGs, X-rays, CTs, and days spent in the hospital for the year pre- and post-enrollment in the program.

Methods
The emergency department has identified approximately 280 patients to date who frequently visit the ED requesting opiate pain medications. A panel of social workers, case managers, and ED administrators reviewed their charts. Upon consensus agreement, the patients were informed that (1) they would undergo referral to an outpatient chronic pain management clinic; (2) they would no longer receive opiate prescriptions from the ED for chronic painful conditions; and (3) an EMR-based notification would inform all ED providers of the patient’s ED pain program referral status. A case management note type entitled “Chronic Pain Program” fires a discern order, activating three Cerner rules. When the patient presents to the ED, the first rule notifies anyone attempting to access the chart of the patient’s referral status to the chronic pain program. To access the chart, the provider must dismiss the notification. A second rule applies a stop sign icon to the tracking board as a visual reminder, and is triggered by the same criteria. The third rule triggers if a provider attempts to order opiates on a patient referred to the program.

Data Analysis
Any patients referred to the program for less than one year were excluded from data analysis to ensure all patients had one year of data pre- and post-enrollment. Wilcoxon signed-rank test was performed for paired medians of non-normal data. Data analysis was performed with RStudio software version 0.98.1073.

Results
243 patients met inclusion criteria and had been referred for at least one year. Median ED visits decreased from 14 per year to 4 (p<0.001). Median CT scans decreased from 2 to 0 (p<0.001), with the outlier patient dropping from 32 to 16 scans. Median X-rays decreased from 5 to 1 (p<0.001), labs performed from 47 to 13 (p<0.001), and ECG’s from 4 to 2 (p<0.001), with the outlier ECG patient decreasing from 158 ECG’s to 49 after enrollment.

Conclusion
This intervention decreased both ED visits and overtesting for opiate seeking patients after they were informed that the ED would no longer provide opiate prescriptions for chronic painful conditions. A major component of the success of this policy is the EMR-based reminders to providers. One downside to this method of implementation is the potential slowdown from too many active rules in the system. Another is provider alert fatigue as ours is not the only discern alert that fires on opening charts.

References
Characteristics of Older Adults’ Adherence of a Wearable Fall Detection Device
Youjeong Kang¹, PhD, Shomir Chaudhuri², PhD, Hilaire J. Thompson¹,², PhD, George Demiris¹,² PhD
¹School of Nursing, University of Washington, Seattle, WA
²School of Medicine, University of Washington, Seattle, WA

Introduction: Falls are the leading cause of fractures and other serious medical conditions related to aging.¹ One in three older adults in the United States experiences falls.² In many cases older adults who have experienced a fall remain on the floor undetected for longer periods of time, exacerbating the negative effects of the fall.³ One informatics strategy to detect falls is the use of sensor technology, such as wearable sensor devices.³ However, little is known about characteristics of older adults who use a fall detection device compared to those who do not.

Objectives: To describe the characteristics of older adults involved in a study exploring the use of wearable fall detection technology as it relates to adherence of wearing the device.

Methods: This is a secondary analysis of a study that explored the use of a wearable fall detection device with a GPS system and automatically detected falls among community dwelling older adults. A total of 18 participants were recruited, and asked to wear the device continuously for a period of four months. As part of their involvement, they were interviewed at baseline, at two and at four months, and were asked to fill out a fall calendar. Additionally, the device provided an online daily log that indicated whether the device had been charged that day and removed from the charger or not. Adherence was measured based on the use of the charger for each day in the study. If the device was removed for more than 20 minutes in a day, it was considered that the participant had used the device. For the purpose of this study, the cutoff percent for adherence was 85% (i.e., participants wore the device on at least 85% of the study days). The cutoff of 85% was determined based on the mean of participants’ wearing time. Participants were classified into the adherent group (≥ 85%) and the non-adherent group (<85%) based on log data. We extracted baseline interview notes. During the baseline interview, participants were asked to state their comfort level with four specific activities: a) using the stairs, b) getting groceries, c) going for walks and d) doing chores around the apartment. Comfort was categorized into two levels (i.e., no, and yes with assistance or yes). We also assessed if subjects wore a personal fall detection device prior to study entry. We utilized the baseline interview notes to identify concerns about the use of a fall detection device, and performed crosstabulation using SPSS version 22 to describe percentages of the comfort level of activities.

Results: All participants had a prior history of falls. Ten participants (56%) were adherent to using a wearable fall detection device. The mean age of adherent participants was 85 years (SD: 9.0), while the mean age of non-adherent individuals was 88 (SD: 5.0). According to baseline interviews, adherent participants were concerned about forgetting to wear the device, while non-adherent participants were concerned about forgetting to charge a device. Fifty percent of those in both the adherent (n=5) and non-adherent (n=4) groups used a fall detection device prior to enrolling in the study. Those in the adherent group reported more comfort among the four activities of daily living, than participants in the non-adherent group [using the stairs (83% vs. 17%); getting groceries (71% vs 29%); going for walks (66% vs 34%); doing chores around their apartments (53% vs 47%)].

Conclusions: Our findings suggest that older adults who are active both inside and outside the home based on responses to the comfort level questions are appropriate target users of a wearable fall detection device with a GPS function. As such devices will only work in those who wear them, future research using a larger sample and a longer study period is warranted to identify factors that contribute to adherence, to increase their use and usefulness.

References

Key words: detection device, fall, older adults

Acknowledgement: This work was supported in part by the NIH National Institute of Nursing Research (NINR) Aging and Informatics Training Program at the University of Washington School of Nursing (Grant Nr. T32NR014833).
Question Types in Online Health Communities
Shaheen Kanthawala¹, MA; Courtland VanDam¹, MA; Barbara Given¹, PhD, RN, FAAN; Joyce Chai¹, PhD; Amber Vermeeesch²; PhD, MSN, FNP-C, RN; Marianne Huebner¹, PhD; John Crowley³, MBA; Jina Huh¹, PhD
¹Michigan State University, East Lansing MI; ²School of Nursing, University of Portland, Portland, OR; ³Alliance Health, Salt Lake City, UT

Introduction
Online health communities and discussion boards can benefit from health experts. However, because of the large scale at which online conversations occur, health experts can find it challenging to moderate posts effectively. Automated identification of questions requiring moderation can streamline this process. To support this automation, we need an annotation codebook, which can help generate training data for the classification of question types.

Among many prior works examining taxonomy of questions ¹ ² ³, Rothwell’s classification of questions⁴ most appropriately addressed the needs of patients’ questions in online health communities. Rothwell categorized questions as (1) fact (asks of something is true and to what extent), (2) value (asks for an evaluation of an idea, object, event or person, which we interpreted as evaluations based on personal experience), or (3) policy (asks if a specific course of action should be undertaken to solve a problem). We built on Rothwell’s model and added (4) social vs. non-social, (5) clinical vs. non-clinical, and (6) question vs. non-questions to address unrelated posts in our dataset (e.g., posts without questions, discussing weekend plans).

Methods
We first examined: (1) What types of questions, based on Rothwell’s model, exist in online health communities? and (2) which question types among them need clinical expertise? Answers to these questions will inform building training dataset for creating automated classification of patient questions in online health communities. We downloaded all questions posted on an online diabetes community from 2007-2014. We selected shorter posts to begin our study with manageable question length: all posts under 100 characters (n=563) and half of the posts with 100-200 characters (n=500). One coder then manually labeled these questions with our classification scheme built on Rothwell’s question types, iteratively checking agreement with another coder.

Results and Discussion
The two shorter question length groups showed similar percentage breakdowns among question types. The types of questions based on Rothwell’s classification as well as those needing clinical attention or not broken down into pie charts in Figure 1. We plan to further examine longer questions to understand whether composition of question types will continue to be similar for longer posts. Our work presented in this poster contributes to developing training data for automatically detecting questions needing health experts in online health communities.

Figure 1: Classification of WebMD diabetes community questions based on Rothwell’s classification

Acknowledgments
This work has been partially funded by NIH 1 K01 LM011980-01.

References
4. Rothwell, J. In mixed company: Communicating in small groups. 2015; Cengage Learning.
aceso (After Cancer Education and Support Operations): a clinical decision support system approach for engaging breast cancer survivors

Akshat Kapoor, M.S.1, Priya Nambisan, Ph.D.1
1University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Introduction
Most modern electronic medical record (EMR) systems today incorporate a clinical decision support system (CDSS) which provides relevant and timely prompts to the physician, based on a set of pre-compiled rules. This technology provides an interactive medium to aid the physician in making clinical decisions. This approach has until now, never been used in consumer health applications intended for patient use. By incorporating this approach in a specialized system (aceso) that supports breast cancer survivors, we explore an alternative means to achieving patient engagement by providing a more interactive online user experience. The system described below is fully developed and is at the beginning stages of testing with end users.

Purpose
The system aims to provide a more, interactive and engaging alternative to patient portals by employing a decision support system approach, in order to support breast cancer survivors.

Problem addressed
Most patient portals in their current state, are a missed opportunity due to their nature of being very generic and aim to serve the entire patient population using a one size fits all approach. There are however special patient groups that could greatly benefit from portals that provide specialized functions. Breast cancer survivors can expect to experience several treatment-related side effects, several weeks after treatment. By employing a clinical decision support systems approach and incorporating feedback in the form of warnings, alerts and reminders for the patient, the system explores making the patient experience more interactive for this group of patients.

System Overview
aceso is an online breast cancer survivorship tool that may be accessed from a variety of web-enabled devices using a simple web browser application. By providing relevant, real time feedback to the patient in the form of warnings, alerts and reminders, aceso actively interacts with the patient. These warnings, alerts and reminders are based on a set of pre-compiled rules, based on breast cancer survivorship guidelines1. Each rule is constructed on the basis of three components: condition (various treatment related side-effects), context (breast cancer related medical history) and action (generation of an alert message or reminder). A benefit of using a set of pre-defined rules in this context is that they are relatively easy to modify and maintain to keep up with changes in guidelines.

Current system state and evaluation
The system has completed design, development and testing phases and IRB approval has been obtained for end user evaluation. Approximately 120 breast cancer survivors are currently being recruited to test and evaluate the system over a period of two months. The system will be evaluated on the basis of its perceived impact on various breast cancer related quality of life indicators, patient-provider communication and patient engagement.

Conclusion
This system has the potential to demonstrate that more personalized and specialized online tools can play a key role in filling the existing gap in the consumer health domain. As patient engagement continues to become a vital component of Meaningful Use Stage 2, healthcare providers should look at alternative means to more effectively engage patients in taking an active role in managing their health in a more interactive manner.

References
OpenMRS and FHIR: The Promise of a Domain Independent API for serving Healthcare Needs Across Underserved Settings

Suranga N. Kasthuriratne, BEng¹, Harsha Kumara, BSc², Burke Mamlin, MD¹,4, Paul G. Biondich, MD, MS³,4
¹School of Informatics and Computing, Indiana University – Purdue University, Indianapolis, IN, USA; ²WSO2 Inc., 20 Palm Grove, Colombo, Sri Lanka; ³Indiana University, Indianapolis, IN, USA; ⁴Regenstrief Institute Inc., Indianapolis, IN, USA

Introduction

The implementation of Electronic Medical Record (EMR) systems across underserved settings is impaired by the lack of technical resources to manage these systems. The enforcement of system specific Application Programming Interfaces (API’s) hinders EMR adoption as they require developers to familiarize themselves with system specific restraints. Another limitation of system specific APIs are that they may not be aligned with any specific interoperability standard, and is challenging to use in support of interoperability between multiple healthcare applications. Given the critical nature of EMR systems there is an acute need to enforce the proper use of standards on implementers of healthcare applications. As an Open Source EMR system that is widely used across underserved settings, OpenMRS has actively felt the need to reduce the learning curve involved with implementing our platform, and to enable standardized data exchange between healthcare applications. We sought to address these limitations by developing a system independent API that could easily be leveraged to promote interoperability. We also sought to design a best approach for a system independent API to work coherently across different versions of a healthcare standard and OpenMRS data model.

Methods

The Fast Health Interoperable Resources (FHIR) [1] standard is a next generation standards framework introduced in response to limitations in HL7 V2 and V3. FHIR combines the best features of HL7’s V2, V3 and Clinical Document Architecture (CDA) messaging standards while leveraging XML, JSON, HTTP and Open Authorization (OAuth). Unlike HL7 V2 and V3, FHIR promotes the use of resources that makes it easier to map them to existing clinical objects within the OpenMRS data model. Upgrading our API to communicate using FHIR would enforce proper standardization of clinical data and reduce the learning curve necessary to use OpenMRS. We sought to develop a FHIR based API that would act as an additional layer above the legacy OpenMRS API, and serve FHIR based resources to users. In addition, we sought to utilize a strategy pattern based approach that allowed the FHIR API to work with different versions of the OpenMRS data model.

Results

We built an add-on FHIR module that would be installed atop the traditional OpenMRS API, and serve as an optional FHIR API to serve FHIR resources to third party users. The initial development phase completed support to export OpenMRS data using a subset of FHIR resources. We also supported a subset of FHIR search parameters based on OpenMRS specific needs. The completed module supports the latest FHIR Draft Specification for Trial Use (DSTU2), and leverages a strategy pattern based handler approach to manage different versions of data objects. Currently, further development work is underway to support the creation and updating OpenMRS data based on FHIR requests. This, as well as additional work to support OAuth security for OpenMRS is expected to be completed by October 2015.

Discussion

By implementing a FHIR based API we resolved the need for implementers to ‘learn’ OpenMRS. Instead of learning a system specific API developers need only learn FHIR, which is appropriate given that it is a standard that is being widely adopted by many EMR systems. We perceive that a standardized API will also enable better interoperability support and help address the challenge of linking together health data fragmented across different EMR systems. During the initial phase, both the FHIR API and the system specific OpenMRS API will be available to consumers of OpenMRS. Following a suitable grace period for users to acclimate themselves with FHIR we plan to deprecate the legacy system specific API in favor of the FHIR based API. We envision that the FHIR API will become the preferred approach for communicating with the OpenMRS API, and be used to serve both internal API calls as well as external requests.

References

Design and Implementation of a Relevant Data Report Tool for Patients Presenting to the Emergency Department with Chest Pain

Chris Katsura, MD, Lynnell McCullough, MD, Frank C. Day, MD MPH.
David Geffen School of Medicine at UCLA, Los Angeles, CA

Abstract
Few clinical decision support tools exist to extract in real-time the pertinent portions of a patient’s voluminous medical record. A patient who presents with chest pain (for example) may or may not have relevant information documented in multiple disparate locations within an electronic health record (EHR). We will create a summary tool that automatically gathers and displays relevant data in a full-report format. We intend to apply this model to other chief complaints and hospital departments.

Introduction
Healthcare providers today are often inundated with massive amounts of stored electronic data. This can be particularly problematic in the Emergency Department, where clinicians constantly face constraints on the amount of time that can be allocated to combing through electronic charts comprised of terabytes of data, potentially missing crucial test results. At our institution, there is no current automated method to compile the data that is relevant to a patient’s chief complaint. Relevant tests are often entered in generically as progress notes or titled under other categories, depending on the generating service. Chest pain is a frequent complaint that comprises five percent of all ED visits, and the patient usually presents with multiple medical conditions and extensive medical records. The likelihood that this complaint represents an acutely life-threatening process may be informed by the results of a wide variety of imaging, procedural, and specialist assessment resources. We hypothesize that inserting a relevant data report into physicians’ standard EHR workflow will result in increased awareness of and review of relevant data sources, and that patients may receive more appropriate workups and dispositions. For practical purposes, we decided to apply our efforts for the initial prototype to patients presenting to the Emergency Department complaining of chest pain, with the ultimate goal of expanding this concept to other common chief complaints (such as abdominal pain, shortness of breath, etc.).

Methods
A multidisciplinary task force comprised of physicians, EHR vendor technical support, and EHR system builders was formed to develop and evaluate a new clinical decision support tool, in the form of an expandable multi-layered side bar that automatically gathers relevant patient data from multiple sources and displays it all in one convenient location. Relevant data can then be easily accessed as a full-report. For patients presenting to the Emergency Department primarily for chest pain, the most recent important cardiac tests (stress, echo, cardiac catheterization, etc.) and cardiology notes are readily accessible at the click of a button. The patient’s medical record will be filtered using categories such as date, exam type, category type, and service to obtain the specific reports. Smart searches for exact terms and synonyms through the physical narrative body of the note will locate inconsistently titled reports and extract pertinent information. In addition, the complete set of all data sources will be validated with colleagues from cardiology and radiology, to ensure that all relevant notes and tests are captured. To determine the effect of the relevant data report tool, we will explore various methods to measure physician work flow, efficiency, door to disposition time, observation admissions, and appropriateness of patient testing.

Conclusion
As a result, the clinician can more quickly and accurately risk stratify a patient complaining of chest pain. Furthermore, this tool will streamline overall workflow, boost efficiency, improve coordination of patient care, and minimize unnecessary cardiac testing. In the future, this prototype model will be implemented across other chief complaints and other medical specialties as a clinical decision support and risk stratification tool.
An Easy-to-Use Clinical Text De-identification Tool for Clinical Scientists: NLM Scrubber

Mehmet Kayaalp, MD, PhD, Allen C. Browne, MS, Zeyno A. Dodd, Ph.D., Pamela Sagan, RN, Clement J. McDonald, MD
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract
Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information prior to their secondary use for clinical research. We have been studying clinical text de-identification for more than a decade and developing NLM Scrubber—it is a tool for every clinical scientist who conducts retrospective research using clinical reports. Although we continuously improve and add new functionalities to it, it is very simple to install and use.

1. Introduction
The Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others; however, manual clinical text de-identification is an arduous task. Furthermore, human annotators alone are usually not as accurate as automatic clinical text de-identification systems. Even though no automatic de-identifier is perfect, they can quickly produce de-identified text, which can then be easily reviewed and verified by the data providers for their de-identification accuracy. If and when the de-identified text needs revisions, the necessary editing is usually minimal.

2. NLM Scrubber
The major downside of commercial de-identifiers is obviously their cost, which many clinical scientists may be unable to afford. Most other de-identifiers found in the literature have been developed for research purposes only and are not available. Besides their no cost to the user, the major advantage of the few freely available de-identifiers is that they can be easily tested, evaluated and verified by independent third parties. The freely available de-identifiers can be further divided into two categories depending on their training data requirements. De-identifiers that require training data impose significant burden on their users demanding a large set of clinical documents annotated in compliance with their prerequisite format.

NLM Scrubber is a freely available automatic clinical text de-identification tool with full support by its developers. Furthermore, it does not impose any annotation requirement on clinical scientists; i.e., no text to be annotated to run the application for producing de-identified clinical reports. Although we continuously add sophisticated functionalities to NLM Scrubber, we strive to keep the user interface as simple as possible so that novice users can operate it easily. NLM Scrubber is a product of several years of studies on clinical text de-identification.1 2 We recently rewrote NLM Scrubber converting it from a pure research product to a consumer product. The user needs to fill out a short form stating mainly where the text files are located in the computer. At this point in time, it can accept only ASCII text reports formatted with proper capitalization; i.e., it would not perform well on all lowercase or all uppercase text.

The system is available on three platforms: Windows, Linux, and Mac OS X. It can be downloaded from http://scrubber.nlm.nih.gov.

Funding and Competing Interests
This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The first author receives royalties from University of Pittsburgh for his contribution to a de-identification project. NLM’s Ethics Office reviewed and approved his appointment.

References
Health Information Technology Evaluation Studies: Trends in Communities and Geography from 2004 to 2014

Marjorie M. Kelley, RN, MS¹, Xiaonan Ji, MS¹,², Gina Torelli, BS³, Po-Yin Yen, RN, PhD¹
¹Dept of Biomedical Informatics, ²Dept of Computer Science & Engineering, ³Dept of Industrial Engineering, The Ohio State University, Columbus, OH

Introduction

Health information technology (HIT) use is increasing, as is evaluation of these technologies. But barriers remain. Using previously outlined definitions of “health information technology (HIT)” and “evaluation research”¹,², we aim to investigate trends in HIT evaluation research between 2004 and 2014 to inform future empirical evaluation research. We hope to inform the research process in an effort to eliminate these barriers in HIT evaluation studies.

Methods

We searched PubMed for journal articles published between 2004 and 2014 for HIT evaluation studies using inclusion and exclusion criteria described by Ammenwerth and de Keizer³. We looked for trends in research groups, author geography, journal name, and the number of HIT evaluation research articles. We also used a novel method to identify research communities. Using ArticleNet⁴, we generated an author network by calculating author similarity for each article pair. This author network identified common authorships among our cohort of articles. We used visualization software to detect dominant communities. Communities with highly connected article nodes shared high author similarity. To better visualize the network, different communities are displayed in different colors with node size representative of author publication activity – larger nodes mean more common author linkages. Contributing authors from the same community, with high collaborative relationships, are detected as “research groups”.

Results

A total of 5,120 articles were found. After applying inclusion/exclusion criteria, 1,624 references remained. Our analysis revealed the largest community (G1) as having the most publications and most common authorship activity. Dr. David Bates, Harvard Medical School, was the most active community member. The largest community (G1) shared significant activity overlap with the 6th largest community (G5). In the second largest community (G2), Dr. Suzanne Bakken, Columbia University, was most active. Interestingly, in the third largest community (G3), authors were active within their own community, but only a few interact with authors from other communities. Three small communities (G4) had no activity with other communities. Studies appeared most frequently in 5 publications: Studies in Health Technology and Informatics, Journal of the American Medical Informatics Association, International Journal of Medical Informatics, BMC Medical Informatics and Decision Making, and Applied Clinical Informatics. These journals published approximately one-third of the articles in our cohort. The number of articles more than doubled in the past 5 years as compared with the previous 6 years. Geographically, the majority of authors were from the U.S. (52.5%), with 29% of these from NY, MA and CA. Authors from NY, MA, CA, PA, MN, IL, TN, OH, MD and TX accounted for just under 65% of all U.S. articles. Interestingly, data imported from PubMed identified USA as country, but left state blank in 5% of cases. Data integrity were concerns.

Conclusion

This preliminary data informs our future HIT research evaluation studies, investigating study methodology, analysis, type of systems, and populations. We hope to identify experts in various HIT evaluation research communities as well as identify potential collaborative opportunities in multi-site HIT evaluation research.

References

Clinico-genomic Decision Support System for Precision Diagnostics and Management
Anuja Kench1, Vandana P. Janeja, PhD1, Amanda Niskar, DrPH2, Yelena Yesha, PhD1, Napthali Rishe, PhD3

1University of Maryland Baltimore County, Baltimore, USA, 2Arthritis Foundation 3Florida International University, Florida, USA

Introduction

The focus of this work is clinico-genomic decision support system for better diagnosing and managing chronic diseases. We propose a clinical decision support system for early diagnosis of chronic diseases (such as arthritis, Diabetes, Cardio Vascular Disease) and better disease management. We propose a generalizable framework, which can accommodate several distinct features coming into the system from clinical, genomic, environmental domains among others. We utilize this framework to discover similar patients and overlaps among patients in a set of these features which has several applications such as: (a) better cohort discovery for clinical trials, (b) better disease management by studying peer group of patients with similar diagnosis and better prognosis, (c) early disease diagnosis by studying similar features in patients with existing diagnosis

Approach:

The overall approach, as shown in the figure, looks at clinical data and performs clinical record extraction. The clinical data consist of the structured medical record, unstructured doctor’s notes, X-rays, lab data among other data components. Some studies have looked at extracting and integrating structured and unstructured data[BM15, G14]. However, our work focuses on integrating multiple types of data from multiple sources with different types of structure. Data is extracted from relational and non-relational forms. Similarly, genomic data is extracted from a large learning database and matched with patient specific genomic data. Databases like VINCI [VA2015] and dbGap[NCBI2015] contain such heterogeneous patients records that can be accessed under agreement. The genomic data is in the form of SNPs extracted from large databases [SW01] and prior clinical studies and patient specific genomic data. Data transformation and preprocessing is performed as new variables are introduced at different stages of the analysis.

The clustering of continuous variables and the categorical variables from clinical and genomic data are combined using a novel algorithm to form clinico-genomic clusters. Data within the clusters is evaluated to find partial or exact patient overlaps. The partial overlaps, especially overlaps above a certain threshold may indicate a high level of similarity in majority of attributes but not in some attributes, which may be of interest to study especially if the overlap is in key clinical variables but not in genomic variables and vice versa. Patients with high overlaps may also be selected as a cohort for clinical trials. The level of similarity can be calibrated to identify various types of overlaps. In addition, multiple diseases can be studied together, to identify potential overlaps between patients from disease groups where no known overlaps exist. This can lead to the discovery of potential links between diseases with no known clinical connections and potentially lead to novel research.

References

Understanding the patient through visualization to improve provider-patient communication in hospitals: Know your patient to personalize your communication

Maher Khelifi, PharmD1, Logan Kendall1, Sonali Mishra1, Andrew Miller, PhD1, Ari Pollack, MD1, Barry Aaronson, MD1,2, Wanda Pratt, PhD3
1University of Washington, Seattle, WA; 2Virginia Mason Medical Center, Seattle, WA

Introduction: The problem of personalized patient-provider communication

Good patient-provider communication plays a major role in hospital quality of care (1). However, good communication requires that providers know not only the patients’ medical information, but also the patients’ communication needs (2). Clinicians tend to have more success when tailoring their messages to best match these needs. Patients have different personalities, values, and preferences; yet, providers have no easy way to understand these differences (3). This lack of knowledge of the patient could negatively influence the patient’s experience and possibly even cause harm. In our study, we aim to understand hospitalized patients’ values and preferences. With this knowledge, we can create a patient profile frame that can be used to visually highlight relevant patient values to providers. These visualizations could quickly give the provider an overview of the patient’s priorities and preferences. With this improved understanding, providers can provide personalized communication.

Methods:
We conducted 28 interviews and observations of adult and child patients and their caregivers across two sites: an adult tertiary care hospital and a children’s hospital. At each site, we conducted 40-60 minute semi-structured interviews with patients and their caregivers. Based on these interviews, we identified a list of possible concerns of the patient during his stay in the hospital. Then, we classified these topics into categories.

Approach: Visualizing Patient Characteristics

Based on prior research and our interviews with hospitalized patients, we identified 7 important categories that the provider has to consider to personalize communication with a patient: Self-efficacy, Privacy, Involvement of caregivers, Details and technical terms, Emotional needs, and Real-time updates (SPIDER). We propose visualizing these characteristics in a radar graph. The graph presents to the provider how much the patient values each of these components. (Figure 1). We hope these visualizations will provide a quick yet thorough view of each patient’s priorities and enable effective and personalized communication. We are conducting more interviews to confirm that our framework is broad enough to cover diverse patient experiences in a hospital. In addition, we seek to confirm that these graphs will be well suited to be incorporated into provider workflow.

Figure 1: How to use the SPIDER Graph?

To identify the patient’s preferences, the patient will be invited to answer a survey and based on his answers we can populate the SPIDER graph. This graph (on the left) shows that for Patient X, the provider has to consider that the caregivers are playing a key role in the patient care. The self-efficacy of the patient is low. So it’s better to communicate with the patient in the presence of his caregiver especially when taking decisions. The patient values knowing about updates and he has a high medical literacy. This means that it is preferable to tell the patient about the last updates about his health care, and it is tolerable to use some technical words with the patient.

References:
Increasing Size of a Health Information Exchange Allows More Accurate Measurement of Early ED Returns

Eugene Kim, MD1, George T Loo, MPA, Tina Lowry, MS1, DrPH1, Bradley D Shy, MD1, Ula Hwang, MD1, Nicholas Genes, MD, PhD1, Cindy F Clesca, MA1, Lynne D Richardson, MD1, Jason S Shapiro, MD1

1The Department of Emergency Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Introduction:
Our group has previously shown that use of data from a health information exchange (HIE) for certain emergency department (ED) quality measures allows more accurate identification of outcomes than site-specific data from individual institutions.1,2 This is because HIE captures data from patient encounters across multiple sites and more accurately reflects the way many patients seek care—across multiple providers. We hypothesized that when two HIEs in the New York metropolitan area merged, that our ability to identify early (72-hour) ED return visits would increase with the larger, post-merger HIE data compared to the smaller, pre-merger HIE data (the latter being a subset of the former). In this study, we compare our ability to identify early ED returns with increasing levels of HIE by measuring 1) site-specific early ED returns with the initial “index” visit and the return visit at the same site, 2) pre-merger HIE-wide early ED returns with the index visit at any of the 10 sites in the original, smaller, pre-merger HIE and the return visit to any of the original 10 sites, 3) post-merger HIE-wide early ED returns with the index visit at any of the 10 original sites, but now with the return visit to any of the 31 sites in the larger, post-merger HIE, and 4) post-merger HIE-wide early ED returns with the index visit at any of the 31 sites in the larger, post-merger HIE, and with the return visit to any of the 31 sites.

Methods:
Data for all ED and inpatient visits from 3/1/09 to 2/28/14 were obtained from Healthix, an HIE in New York City. Data were cleaned and de-identified in accordance with Healthix and New York State policy and analyses were performed to identify the number of early ED return visits for each of the 4 analyses described above. Only ED visits were used for the index visit, but either ED or inpatient visits were used for the returns.

Results:
A total of 12,669,657 encounters across the larger post-merger HIE were used in the analysis, including 6,352,829 encounters from the smaller, pre-merger HIE subset. Site-specific early ED return rates (early ED return visits divided by all ED visits) ranged from 0.7% to 8.8% across all 31 sites (median 2.7%). Pre-merger HIE-wide data with the index visit at any one of the 10 pre-merger HIE sites and the early return visits at any of 10 sites across the HIE yielded early ED return rates ranging from 3.7% to 15.4% (median 5.65%). Post-merger HIE-wide data with the index visits at any one of the original 10 pre-merger sites and the early return visits at any of 31 sites across the new, larger HIE yielded early ED return rates ranging from 3.8% to 15.5% (median 5.7%). Post-merger HIE-wide data with the index visits at any one of the 31 sites and the early return visits at any of the 31 sites across the new, larger HIE yielded early ED return rates ranging from 1.8% to 15.5% (median 4.5%). The average total incremental increase in our ability to identify early ED returns increased by 66% going from site-specific data to measurement across all 10 sites in the pre-merger smaller HIE; and by 85% going from site-specific data to measurement across all 31 sites in the post-merger larger HIE.

Conclusions:
This analysis shows that with increasing data size of an HIE there was an incremental benefit in our application of a common ED quality measure. Compared with site-specific data, the average ability to identify early ED returns increased by 66% in the smaller HIE (10 sites), and by 85% in the larger HIE (31 sites). As local and regional HIEs merge and consolidate into larger nodes of aggregated clinical data, toward the eventual goals of state-wide and nationwide HIE, our ability to more accurately apply clinical quality measures should also increase. Further analysis should be done to determine if similar incremental benefits with larger HIEs are found for other quality measures such as 30-day hospital readmissions.

Usability of mobile apps for radiology diagnostic decision-making

Min Kim, PhD1, 2, Michael R. Aro MD3, Kraig J. Lage MD3, Kevin L. Ingalls MD3, Vivek Sindhwani MD3, Mia K. Markey, PhD4
1Department of Health Management and Informatics, University of Missouri2; University of Missouri Informatics Institute; 3Department of Radiology, University of Missouri, Columbia, MO; 4Department of Biomedical Engineering, The University of Texas, Austin, TX

Abstract

We evaluated the usability of eight iOS mobile apps currently available in radiology to support training in diagnostic decision-making. Lab-based usability tests using video analyses with the triangular method approach were conducted to determine usability issues. The methods include video analysis, system usability scale (SUS), and a debriefing session with participants. Six staff radiology physicians completed a typical set of tasks, using think-aloud strategy. Thirty-five unique usability issues were identified and 18 improvements were suggested.

Introduction

Radiology physicians are increasingly using mobile devices in image analysis for decision making. However, very few of the healthcare apps available on the market have been tested with respect to their usability. The purpose of this study was to identify and measure usability issues and collect suggested usability improvements from radiologists.

Methods

A multi-step review process was used to identify mobile apps that assist in education and training of radiological diagnostic decision-making processes in two major stores (Google Play and iTunes). From 381 apps that were initially identified, eight iOS apps that deal with a variety of radiology modalities were selected for usability evaluation. Three apps were available in both stores: (1) Case Review, (2) AART Ultrasound Cards Lite, (3) Dexnote, (4) Brain MRI Sectional Walker, (5) Diagnostic Radiology - Dynamic Approach to Abdominal Radiology, (6) MED imaging Case, (7) Radiology Assistant, (8) Radiology Head. The apps were installed on an iPad for data collection. Six staff radiology physicians performed scenario-based tasks in a lab-based setting. A mobile observation device was used to record screen activities and hand gestures, which allowed comprehensive capture of the human computer interaction.

Results

The SUS (0-100) score illustrated that physicians ranked the app’s usability at a mean of 67 (high marginal) and the score ranged from 41 (AART) to 76 (Diagnostic Radiology). Task analysis revealed 35 unique usability issues. The most frequent issue observed was inefficient navigation; most physicians had to spend extra time testing inactive functions due to non-intuitive or misleading labels, which was consistent with physicians self-reported 18 unique improvements.

Example recommendations were matched according to the usability principles suggested by mHIMSS (Table 1).

Table 1 Examples of suggested usability improvements and related mobile usability principles from mHIMSS.

<table>
<thead>
<tr>
<th>Suggested usability improvement</th>
<th>Usability principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Clear indication of buttons or areas to tap to enter a submenu; keep original menu item visible when opening submenus; arrow keys for scrolling in a stack in addition to finger drag”</td>
<td>Minimize cognitive overload, Preservation of context, Efficient Interaction</td>
</tr>
<tr>
<td>“Stick to standardized iPad multi-touch gestures. Scrolling by dragging a separate grid is not intuitive.”</td>
<td>Consistency, Efficient Interaction</td>
</tr>
<tr>
<td>“I always want to touch the image to scroll on it; I like programs that are simple to use, no complexity”</td>
<td>Efficient Interaction, Simplicity</td>
</tr>
<tr>
<td>“Things which require long clicks/double clicks are less intuitive. Small universally understood signs help for navigating through the app.”</td>
<td>Efficient Interaction, Consistency</td>
</tr>
</tbody>
</table>

Conclusion

This study (1) demonstrates the good example of usability evaluation, and (2) suggests potential improvements in the development of medical mobile apps in radiology education and training. Currently, there are no systematic and standard usability guidelines for mobile medical apps. Consequently, the approval process of the applications does not require a measure of usability. While SUS scores indicate the usability of the apps as acceptable, physicians reported some issues in this small pilot study. This suggests the urgent need of systematic usability evaluation for potentially improving user acceptance and achieving educational effectiveness. Further research is warranted involving different platforms (Google Play, Windows, BlackBerry) and varying sizes of mobile devices (smartphone, phablet) to achieve generalizability.
Improving Detection of Reasons Not to Take a Medication by Leveraging Medication Prescription Status

Youngjun Kim1,3, Jennifer Garvin, PhD, MBA2,3, Julia Heavirland3, Stéphane M. Meystre, MD, PhD2,3
1School of Computing, 2Department of Biomedical Informatics, University of Utah; 3VA Health Care System, Salt Lake City, Utah

Abstract: When automatically assessing heart failure treatment quality measures, identifying mentions of reasons not to take heart failure treatment medications in a patient electronic health record is important. We created machine learning-based sequential taggers to extract reasons not to take heart failure medications based on lexical features. When adding the medication prescription status as feature, performance improved with 41% F1-measure.

Introduction: Extraction of Heart Failure (HF) treatment measures from clinical records is a crucial step to evaluate whether HF patients are provided recommended care. The CHIEF (Congestive Heart Failure Information Extraction Framework) application was developed to extract left ventricular ejection fraction (EF) mentions and values, Angiotensin Converting Enzyme Inhibitor (ACEi) and Angiotensin II Receptor Blocker (ARB) medications, and reasons not to take these medications (RNM). The performance of EF and medication detection [1-2] was quite good. However, the extraction of RNM is still considered a difficult task because RNM are only rarely found in clinical records and usually in less structured formats.

Methods: As part of the Veterans Health Administration HMP (Health Management Platform) project, we developed a reference standard of 1,655 clinical notes manually annotated by domain experts. Among these notes, 790 were used for training and 865 for testing. RNM was less frequently mentioned than EF or medication information with only 204 mentions in the training set and 241 mentions in the test set. As a baseline, we created a rule-based system using a dictionary lookup approach with a manually-built list of RNM terms. We then built machine learning-based sequential taggers to detect RNM based on lexical features. As specified in our annotation guideline, a RNM was annotated most frequently when the patient was taken off an ACEi or ARB medication. This observation motivated us to also use the medication prescription status to improve performance of RNM detection.

Results: For the dictionary lookup, we compiled a list of RNM terms from our annotation guideline and existing standard terminologies. This rule-based system only reached low performance with a 5.3% F1-measure on test data. The machine learning-based sequential tagger models were trained with 135 notes containing RNM (about 17% of training data). When using features including medication prescription status as classified by one of the CHIEF module [2], our sequential tagger increased precision to 42.3% and the F1-measure reached 41%.

<table>
<thead>
<tr>
<th>Method</th>
<th>Medication prescription status feature</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary matching</td>
<td>N/A</td>
<td>31.5</td>
<td>2.9</td>
<td>5.3</td>
</tr>
<tr>
<td>Sequence Tagging</td>
<td>Without</td>
<td>41.5</td>
<td>23.9</td>
<td>30.4</td>
</tr>
<tr>
<td>Sequence Tagging</td>
<td>With</td>
<td>39.8</td>
<td>42.3</td>
<td>41.0</td>
</tr>
</tbody>
</table>

Table. Reason not to take medications detection results (Percentage)

Conclusion: This study shows that the sequential modeling outperforms rule-based system in detecting reasons patients are not prescribed evidence-based medication in this study. Our sequential tagger, incorporating medication prescription status that can be satisfactorily classified, allowed for improved performance.

Acknowledgments: This research was supported by VA HSR&D IBE 09-069, HIR 08-374, HIR 09-007, and funding from the VA Office of Informatics and Analytics. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or those of the University of Utah.

References:
Title: Analysis of the Great Divide Between Cardiovascular Risk & Health Scores

Bobbie Kite, PhD, MHS\textsuperscript{a}, Tasneem Motiwala, PhD\textsuperscript{a}, Philip Payne, PhD, FACMI\textsuperscript{a}

\textsuperscript{a}Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA

Problem:
Cardiovascular (CV) assessment may be performed through risk and health scores. CV risk scores are commonly used in a clinical setting to assess risk of cardiovascular disease (CVD). Cardiovascular health (CVH) scores are typically used in a consumer setting to assess individual lifestyle factors. CV risk and health scores are derived from disparate sources of data for different audiences. Variance in these factors creates voids between patient and clinician perceptions. These divides in turn, may translate into usability issues. Articulating this divide is the first step towards creating bridges and solutions.

Methods:
Using a population of approximately 35,000 members from The Ohio State University Health Plan, we characterized the CVH score along with Framingham 10yr Cardiac Disease Risk Score using clinical measurement guidelines and risk scoring algorithms for those with available data. A score was calculated for each eligible member per measure and population health metrics were formulated.

Results:
We were able to characterize 33,821 and 34,810 members for CVH in 2013 and 2014 respectively; and 23,858 and 27,725 for 10yr CV heart disease risk in 2013 and 2014 respectively. This represents a 9% growth in the portion of the population we were able to characterize for CVD risk in 2014. The average mean score for CVH in this population was 61.96% for 2013 and 62.38% for 2014; and the average CVD risk was 2.96% and 2.94% respectively. Where 18\% of the population was characterized with poor CVH (49\% or below), only 1.9\% of the population was screened with a high 10yr risk of CVD (>20\%). Conversely, only 12\% of the population was characterized with ideal (80\% or above) CVH and 88\% was screened with a low 10yr risk of CVD (<10\%).

Conclusions:
The need for a domain analysis of the divide between CV risk and health is necessary because consumers and clinicians utilize informatics based assessment tools receiving conflicting information. This situation is compounded by informatics issues such as missing data and variable granularity along with a lack of behavioral models within this domain. As we seek to answer research questions requiring usability of assessment tools, we can leverage informatics to bridge these divides. Next steps include a systematic review of clinician and patient motivations, related informatics, and appropriate applicable behavioral models.
The Scalable Collaborative Infrastructure for a Learning Health System

Jeffrey G. Klann, PhD1,2,3, Marc Natter, MD1,4, Doug MacFadden, MS1, Sarah Weiler, PhD1, Kenneth Mandl, MD, MPH1,4, Shawn N. Murphy, MD, PhD1,2,3

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

Background
The Scalable Collaborative Infrastructure for a Learning Health System (SCILHS, pronounced “skills”) is a growing network of health centers across the United States, presently covering over 8 million patients at eleven sites. [1] SCILHS is a Clinical Data Research Network (CDRN) in the Patient-Centered Outcomes Research Institute’s PCORnet, a national effort to instantiate a ‘network of networks’ that supports large-scale comparative effectiveness research. [2] SCILHS has adopted the PCORnet Common Data Model (CDM) as its foundation for interoperable data exchange.

SCILHS uses Informatics for Integrating Biology and the Bedside (i2b2) as its technical backbone. i2b2 is an open-source clinical data warehousing and analytics platform funded by the National Institutes of Health. It is used at over 100 sites nationwide, including several CDRNs and the National Center for Advancing Translational Sciences (NCATS) Accrual to Clinical Trials (ACT) network. i2b2 supports live distributed queries through the Shared Health Research Informatics Network (SHRINE) platform, which SCILHS has adopted and enhanced.

System Description
SCILHS is enabling patient-centric clinical research through:

• **Live cohort-finding queries across all its sites with real-time results.** We developed an i2b2 ontology that represents the PCORnet CDM, and we created a mapping methodology using existing i2b2 components that allows sites to support this common ontology without changing their underlying data. By developing a mapping to this ontology (the SCILHS CDM), sites can participate in the SCILHS SHRINE, which provides a graphical query tool to develop and execute prep-to-research queries across all sites in real time. We have made our ontology publicly available, and portions of it are being used at several other CDRNs. [3]

• **Incorporation of patient reported outcomes data with clinical data.** An emerging need in clinical research informatics is the inclusion of patient reported outcome (PRO) information in research data, such as subjective evaluation of quality of life with chronic disease. The mySCILHS platform provides a framework to gather and integrate PROs with i2b2. Patient survey questions are defined in the popular Research Electronic Data Capture (REDCap) tool, and a combination of REDCap and a commercial telephony platform (CXP, Aspect Software, Inc) are used to survey patients. The PROs gathered via telephony from patients assigned to REDCap surveys are delivered to an i2b2 Loader service that stores the data in an i2b2 repository. We have implemented surveys that focus on quality of life and symptom burden for these diseases in conjunction with clinical experts.

• **Interoperability with the PCORnet Distributed Research Network for nationwide data analytics.** We developed a tool that transforms data in an i2b2 repository into the PCORnet CDM tables that are used for PCORnet DRN queries. This tool relies on the mappings developed to support the SCILHS CDM, so an automated process enables DRN interoperability at compliant sites. Seven of our sites have transformed their most complete data into CDM tables and have successfully responded to DRN queries made through the PopMedNet query tool.

Discussion
SCILHS is paving the way for large-scale patient-centric comparative effectiveness research, building on 15 years of informatics innovations. It offers a common data model that can be implemented without changing underlying data, real-time query results, interoperability with the PCORnet Distributed Research Network, and a telephony framework for gathering and incorporating patient reported outcome data.

References


Mixed-Methods Study of Risk Communication in a Patient Decision Aid

Krystal A. Klein, PhD, Lindsey Watson, BS, Latha R. Kalaga, MS, MBI, Karen B. Eden, PhD
Oregon Health & Science University, Portland, Oregon, USA

Introduction

Overall, patients who used patient decision aids—evidence based tools designed to educate patients about their options concerning a specific medical decision—in conjunction with typical care have superior knowledge and risk comprehension scores relative to patients who receive care as usual.¹ The details of how to communicate risk to patients within a patient decision aid must be carefully considered on a case-by-case basis. Currently, risk communication is as much an art as a science, despite a growing literature that maps the effects of risk formats, viewer characteristics, and chart types on outcomes such as perception of risk levels, gist or exact recall of risk statistics, and decision outcomes (i.e., application of risk knowledge). This mixed-methods study recorded the impressions and reactions of rural patients, in their own words, to quantitative risk information in Mammopad², a mobile app patient decision aid for women in their forties to help them decide at what age to begin having routine screening mammograms. Through this study, we aimed to gain understanding of the role and perceived importance that risk information in a decision aid has in a medical decision-making experience. Simultaneously, we investigated the effectiveness of Mammopad’s specific risk communication graphics, by probing recall of key statistics and success at answering a word problem that required application of the presented risk information.

Methods

Risk presentation quality was evaluated in three pieces: 1) A risk scenario question where participants answered a word problem with their quantitative estimate of the positive predictive value of screening mammograms for women in their forties (administered before and after using the decision app); 2) A thematic analysis of transcripts from interviews with a subsample of Mammopad participants about what they found valuable about quantitative risk information; and 3) A thematic qualitative analysis of transcripts from the same interviews about the interpretation of risk communication diagrams, including misperceptions/misinterpretations of data.

Results

Estimates of positive predictive value of mammography by participants (n=71) were more accurate after using the aid than they guessed before using it, although that was partly due to confusion between different risk statistics presented in the decision aid. Twenty-one participants enrolled in the interview phase: twelve participated in an interview session within an hour of using Mammopad, and nine others were recalled weeks or months after first using Mammopad. Our analysis discovered the following negative themes concerning numeric risk presentation: 1) lack of gradations in perception of uncertainty based on numbers, 2) numbers are sometimes provided as explanation, instead of a tool for explaining, 3) skepticism about the value of numeric information in light of forgetting, and 4) confusion about different statistics. Positive themes were that the women: 1) valued grounding in real-world groups, 2) valued a connection to medical research, and 3) valued transparent enumeration of outcomes. A simple breast cancer incidence pictograph was well comprehended by participants, and recall for risk information was good shortly after using the aid (but quite poor after a latency of weeks or months). A more complicated flow-chart graphic was miscomprehended in several different ways that we present here.

Conclusion

The numerical risk graphics in Mammopad were well received and informative, although not memorable on a long-term basis. It may be helpful to administer Mammopad at multiple times during a patient’s forties to refresh her memory. The more complicated of the two risk graphics in Mammopad was less often completely read or comprehended by participants, but cognitive tools such as animation could ease the cognitive burden of comprehending this information and improve understanding.

References


Interactive Voice Response Technology: Promises and Pitfalls in Facilitating Patient-Reported Monitoring for Adverse Drug Reactions

Elissa V. Klinger, MS1; Alejandra Salazar, PharmD1; Jeffrey Medoff, BS1, Mary Amato, PharmD1,3, Patricia Dykes PhD1, Jennifer S. Haas, MD1,2; David W. Bates, MD1,2; Gordon D. Schiff, MD1,2

1 Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA; 2 Harvard Medical School, Boston MA; 3 Massachusetts College of Pharmacy and Health Sciences, Boston MA

Introduction: We developed a patient-reported, interactive voice response system (IVRS) to actively monitor the safety and effectiveness of treatment for patients taking FDA-approved medications for one of four common chronic conditions (diabetes, hypertension, insomnia, depression), with real-time support by a pharmacist. We present an analysis of our experience highlighting promises & pitfalls associated with this technology and offer strategies for other healthcare organizations that might seek to create a similar patient-reported outcome platform.

Methods: We identified and described 6 major activities in the deployment of an advanced IVRS platform: (1) developed an algorithm to identify primary care patients who had newly started a target medication; (2) created an IVRS-delivered data collection tool that included consent, patient ID, broad /targeted symptom screening with branching logic, and live-transfer to a pharmacist; (3) worked with in-house telecommunications to ensure appropriate caller IDs & patient dial-back routing; (4) created workflows to capture IVRS patient-reported data to pre-populate the pharmacist database for symptom validation and intervention documentation; (5) fired alerts off of positive responses to queries about symptoms; and, (6) developed an in-house customer-service platform to provide support/assistance with calls and an opportunity to opt-out in real-time.

Results: Of 4,847 eligible patients called by IVRS, 609 (12.6%) participated & provided data; 309 reported positive symptoms DQGZHUHWUDQVIHUUHGWRWKHSKDUPDFLVW:KLOHWKLVGHPRUQWU PDWHV³VXFFHVV´LQWKHRSHUDELOLWDQG, the platform we describe is not without challenges, for both end-users and the healthcare system into which it is embedded. Of 25,670 potentially eligible patients identified in our EHR, 11,945 (46.5%) were eligible true new target drug starts, illustrating the challenge of correctly identifying patients in an EHR who accurately qualified for the intervention (e.g., many apparent new starts actually started outside our system). The rigidity of a structured & inflexible IVRS was confusing and frustrating. Patients had privacy concerns; although sent a letter in advance, some objected to seemingly unannounced calls and reported that they would have preferred knowing content in advance. Future interventions might inform patients about planned calls at the time of their clinic visit, thereby allowing opt-in/out and creating the expectation of the call as part of or an adjuvant to their care. Overly-complex branching logic in IVRS data collection tools made the calls long and created misunderstanding; patients are more likely to hang up the longer the call continues without resolution, thus affecting data quality and subsequent clinical action. Pre-loading outbound calls with specific patient data (without compromising confidentiality, should someone other than the patient answer the phone) is a step toward personalization, and underscores to need for accurate algorithms to identify appropriate patients. Email alerts that fired based on patient-inputted information were an important step in notifying team members about actionable data in real-time; however, the alerting system was undermined by delays due to web traffic and firewall “hops.” Caller IDs from the IVRS server should be “replaced” so that the hospital/health system name appears; any patient call-backs must route directly to a live person or a mailbox checked routinely, a requirement that we achieved only midway through the project. Pharmacist interventions included filing a templated note manually in the EHR. For patients who did not talk to the pharmacist but reported a symptom, data did not enter the EHR; ideally all patient-reported data would interface seamlessly. Providing real-time customer service is an essential component to the success of any automated call platform. The option to talk to a live person immediately at any point during the automated call should be considered for any IVRS intervention in which patients may require further assistance. Finally, the comparative effectiveness of IVRS technology vs. other forms of efficient/automated synchronous and asynchronous communication to capture patient-reported outcomes needs additional research, particularly in light of the myriad issues that IVRS raised, and may make other approaches better to try.
Text Mining of Patient Demographics and Diagnoses from Psychiatric Assessments

Eric J. Klosterman, M.S.¹,², Rashmi Prasad, PhD²

¹Rogers Memorial Hospital, Oconomowoc, WI; ²University of Wisconsin-Milwaukee, Milwaukee, WI

Introduction: The present work contributes to the currently understudied application of text mining from narrative texts of psychiatric assessments, by exploring keyword-based pattern-matching and Natural Language Processing (NLP) methods for automatic extraction of patient demographics (age, gender, marital status, and education level) and Axis I admission diagnoses, which are all also automatically assigned a code. Applications of such systems include determining patient outcomes and eligibility for clinical trials. While there has been research on the automatic extraction of medical diagnoses [1], none yet has dealt with psychiatric diagnoses, and work on extraction of patient demographics are limited in both depth and scope [2].

Methods: 200 manually de-identified admission psychiatric assessments of patients from a mental health hospital (Rogers Memorial Hospital (Oconomowoc, WI)) were manually annotated (double blind) by two annotators with extensive experience in reviewing psychiatric assessments for coding purposes. Annotation involved code assignment to each document, and marking text spans indicative of each code. Separate algorithms were developed for each information category over 110 documents used as the development (dev.) set, with the remaining 90 documents used as the test set. Keyword-based pattern matching rules were first applied, and then augmented with NLP-based rules (defined over constituent trees and typed dependencies obtained with the Stanford Parser [3]), to improve performance. Results: Evaluation, over the dev. and test sets, used two metrics of code accuracy (No. of correctly coded documents/No. of total documents) and text accuracy (No. of documents coded with positively matched text/No. of total documents). We found that age and gender appeared consistently in semi-structured document sections and were almost entirely identifiable using patterns, with code and text accuracies above 98% in both dev. and test sets. Detection of (primary) diagnoses text spans and their coding was also done with a pattern-based approach, but the performance dropped from 99% in the dev. set to 92% in the test set. Some errors resulted from patterns but others occurred primarily because the UMLS Metathesaurus DSM-IV vocabulary used to detect diagnoses strings was not comprehensive enough to account for the wide variation in the observed terms. Education level and marital status, located in the unstructured portion of the narrative, were the most challenging because of the wide variation in their linguistic expression. Results (Table 1) show that structural rules improved both code and text accuracy, although there was a more significant drop in performance from the dev. set to the test set for education level than for marital status, indicating the need for better generalization of the rules for education level. Results for the pattern-based approach for marital status indicate that better performance on text span identification does not correlate with better code accuracy since spans can contain negations and modality that affect the code assignment.

Conclusion: A text mining method combining keyword-based patterns and rules designed over sentential constituency and dependency structures can achieve high accuracies for extracting patient demographics and admission diagnoses from psychiatric assessments. Limitations arise from errors in pre-processing components, inadequacy of terminological resources for term coverage, and generality of the rules themselves. This study makes a novel contribution to text mining work in the psychiatric domain, with promising initial results.

<table>
<thead>
<tr>
<th></th>
<th>Education Level</th>
<th>Marital Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev. set</td>
<td>Test set</td>
</tr>
<tr>
<td>Pattern</td>
<td>88.2%</td>
<td>74.4%</td>
</tr>
<tr>
<td>Pattern + NLP</td>
<td>92.7%</td>
<td>77.8%</td>
</tr>
</tbody>
</table>

Table 1: Results for Education Level and Marital Status

References:
Abstract
In this research we document the challenges and lessons learned during the development of the BigMouth dental data repository. BigMouth uses the i2b2 framework and contains EHR data from six dental schools. We faced various challenges including: a) issues related to the multi-institutional nature of the project, b) data modeling and integration issues and c) infrastructure related issues. BigMouth DDR is one of a kind data repository in dentistry that could serve as an important resource for research.

Introduction
BigMouth Dental Data Repository (DDR) is an inter-school dental data repository that contains data from the axiUrn Electronic Health Record (EHR) used at six dental schools which are part of the Consortium for Oral Health Research and Informatics (COHRI). BigMouth provides researchers with a single large database to advance evidence based dentistry. BigMouth was built on the Informatics for Integrating Biology and the Bedside (i2b2) data warehousing platform. Extracting and loading data from six different institutions present unique challenges. The purpose of this research is to report on the different challenges faced during the development and the approaches undertaken to address these issues.

Method
A set of configurable scripts that extract data from the axiUrn EHR were created during the start of this project. Each school was provided with these scripts during the project release. The schools ran these scripts against their EHR instance to produce a set of Comma Separated Values (CSV) files that corresponded to the database tables in i2b2. These files were securely transferred to UTHealth and loaded into BigMouth using a data integration tool called Talend. Once completed, users from the participating schools can securely authenticate into BigMouth using a federated identity management solution provided by InCommon. Each user has access to the complete dataset extracted from their school and also to an integrated terminology system that allows users to query data across all schools that contribute data to BigMouth. The integrated terminology hides the source information providing the schools full control over their data.

Challenges and Solutions
Collaborating with multiple institutions came with various logistic challenges such as communication delays, educating remote teams about the settings, execution of data extraction scripts, debugging, data cleaning and terminology mapping. In order to minimize the impact of these issues on the development of the repository, we adopted a centralized data integration approach where data are refreshed on a half yearly basis providing us the opportunity to perform data quality checks. These quality checks includes manual sanity check using a checklist and scripts that compare new data with the previous release to ensure no data are corrupted or lost.

A common problem in data integration is the representational heterogeneity of data. Data from different sites contain a) Content differences – granularity or level of information captured b) Naming differences – use of synonyms etc. c) Semantic differences – meaning of same term varies based on the context in which it is used. As a solution to these problems, two views were developed that can be seen by the users in the user interface. A user can either query for data at their own school by using the local terminology that they are familiar with or query data across all sites using an integrated terminology called the COHRI terminology. Each folder in the COHRI terminology was either adopted from standard terminologies or developed from the scratch. Standard terminologies were used for diagnosis and procedure information. For the rest of the concepts like demographics, periodontal and forms data, a standardized terminology system was developed with the help of domain experts at COHRI. In fact, this effort encouraged COHRI members to collaborate and create standard forms for medical history, dental history and caries risk assessment. As of March 2015, BigMouth contains data on over 2 million patients from six institutions. The infrastructure is being constantly evaluated, tested and upgraded based on the performance needs.

Conclusion
BigMouth contains EHR data from six institutions that are being used for research and advancement of evidence based dentistry. The basic data extraction and loading processes have been refined since the initial development of the project. This has resulted in the creation of a scalable infrastructure that can provide a robust platform for future addition of data from dental institutions.

References
A Novel Visualization for Rapid Summarization of Patient History: Application to Cirrhosis

Jejo Koola\textsubscript{1,2}, Samuel B. Ho\textsubscript{3,4}, Michael E. Matheny\textsubscript{1,2}

\textsuperscript{1}Tennessee Valley Healthcare System, Dept. of Veterans Affairs, Nashville, TN
\textsuperscript{2}Vanderbilt University, Nashville, TN
\textsuperscript{3}University of California at San Diego Medical Center
\textsuperscript{4}Dept. of Veterans Affairs, San Diego, CA

BACKGROUND: Patient phenotype identification serves as an essential step for clinical trial recruitment, outcome prediction, survival analysis, and other retrospective and prospective research studies. Physicians diagnose chronic diseases and interpret severity of illness by interpreting the large volume of clinical information available in the patient’s electronic health record (EHR).

OBJECTIVE: We sought to develop a visualization to succinctly summarize a patient’s clinical history relevant to advanced liver disease progression to support rapid clinical interpretation.

METHODS: A radial layout with the time axis along the circumference was developed to represent the evolution of key clinical information values. Concentric rings within the visualization model represent input features. Continuous variables, such as laboratory tests, were mapped from normal to abnormal in the direction of expected liver disease using a color scale from green to yellow to red. Categorical information, such as inpatient admissions, outpatient visits, and radiology studies were mapped to discrete markers. The D3 JavaScript engine was used because one can develop an interactive user interface that allows drill-down and changes in representation by the user. As a pilot, we analyzed a retrospective cohort of patients receiving care at the Dept. of Veterans Affairs between 2005 and 2013. We selected patients with known risk factors for cirrhosis (based on ICD 9 code) who developed advanced liver disease during the study time period. We extracted and modeled 7 features: creatinine, albumin, Model for End-Stage Liver Disease (MELD) score, bilirubin, platelet count, prothrombin time, and inpatient admission dates.

RESULTS: Figure 1 represents the results of the visualization tool for one of the sample’s selected patients. The patient evinces a pre-cirrhosis phase between 2005 and 2007 where liver function is normal. During this time the patient’s main risk factor for developing liver disease was Hepatitis C. In late 2007, the patient showed the first signs of liver dysfunction, and had an inpatient admission for the purposes of a liver biopsy. The patient then has a stable quiescent phase between 2008 and 2009. Starting from early 2009 the patient enters a progressive decline in liver function until he expired from cirrhosis complications in mid-2010. The individual rings allow for clear determination of these distinct clinical situations.

DISCUSSION: Cirrhosis is a complex, sometimes progressive, fatal disease. Though liver biopsy serves as the gold standard, more commonly, physicians use laboratory values, clinical history, and radiology to make the diagnosis. Overall, this pilot work highlights that thoughtful visualization allows the display of complex clinical information representations over an extended time period. Healthcare providers and secondary data users can quickly identify these disease state changes for use in clinical care as well as annotation and chart review purposes.
**Clostridium Difficile Repeat Ordering Difficult to Control without CDS**

Vishnoo Charan Reddy Kothapeta, MBBS, MBA¹, Alfred Caligiuri, PA, MBA¹, Kevin Bock, MD, MHCDS¹, Michael Oppenheim, MD¹, Eytan Behiri, MD¹

¹North Shore-LIJ Health System, Lake Success, NY

**Introduction**

CDC-approved latest guidelines discourage repeat Polymerase chain reaction (PCR) testing in patients suspected of or diagnosed with Clostridium difficile infection (CDI). However, at North Shore LIJ Health System, there had been a significant number of repeat orders for C. difficile Toxin by PCR. System-wide policy is to place any patient with a C. difficile Toxin by PCR order in private bed isolation leading to resource wastage. The Computerized Provider Order Entry (CPOE) system with embedded Clinical Decision Support (CDS) has provided an opportunity to improve adherence to evidence-based knowledge guidelines. Since duplicate order checking is not specific enough in this clinical scenario, hard stops were used for blocking repeat orders, as they have shown to be more effective than “soft” alerts in preventing adverse events.

**Method**

Intervention: To reduce repeat orders, a Medical Logic Module (MLM) was designed and implemented at hospitals which use Allscripts’ Sunrise Clinical Manager as the Electronic Health Record (EHR) system. A hard stop (blocking the order) was implemented to block the provider from placing a new order in the CPOE system; if the patient has a recent result of either a negative within the previous 3 days or a positive within the previous 3 weeks. The provider is shown an alert message to inform that a recent result is available. Additionally, there was education and outreach to inform providers of these new policies.

Study Design, Setting: This is an uncontrolled interventional study analyzing effects of the hard stop on C. difficile orders during the following periods: four months prior (7/15/2014 to 11/11/2014) to implementation, four months after (11/12/2014 to 03/11/2015) implementation and the previous year (11/12/2013 to 03/11/2014) to compare season variations. Medicare/Medicaid reimbursement ($50.27) was used to calculate lab savings. Excess cost of a private bed ($148.75) was calculated using pricing information from Rochester General Health System, NY.

Data Collection and Statistical Analysis: Data related to the number of orders placed, orders blocked and alerts displayed were obtained by SQL queries and exported to Microsoft Excel to conduct an unpaired t-test analysis.

**Results, Lessons Learned, Key findings**

A two-tail p-value of less than 0.05 was considered significant. Statistically significant decrease is observed when post intervention data (1174 orders) was compared with pre intervention 1528 orders (p-value = 0.003) and with last year’s 1863 orders (p-value < 0.001). Prior to MLM implementation, 125 repeat orders were placed including 6 paper orders. After implementation, there were still 33 repeat orders and the number of block-able CPOE orders dropped to 20 (orders outside the scope of current MLM).

During the study period, 230 alerts were displayed to block 117 orders of which 65 orders were for patients who had a previous negative result. This is a $9,668.75 reduction in resource wastage by not assigning private beds; assuming that the only change in treatment modality is private bed isolation. Additionally, laboratory resource utilization was reduced by $5,881.59 which yields a 4-month total of $15,550.34 resource improvement for the health system.

**Discussion – Key conclusions**

The MLM intervention achieved a statistically significant decrease in the number of orders by creating a full hard stop. In an ideal world, a provider should gain immediate value from an alert displayed. Hence, to reduce clicks or provider effort to review the referenced result, an improvement to the existing alert would be to show date and time of the result as well. While education and outreach can create a change in a provider behavior pattern, a CDS forces providers to adhere to evidence based medicine and the return on investment is larger. CDS can be effective but if overdone, can lead to a clinically important treatment delay which would gravely affect quality of care. The key to CDS is moderation and effective use with the right balance as to not affect clinical outcomes.

**References**

A Lifecycle Management Solution to Manage Mississippi’s Data Lake and Big Data Analytics Platform

Denise D. Krause, Ph.D.
University of Mississippi Medical Center;
Office of Mississippi Physician Workforce

Background. Mississippi leads the nation in poor health outcomes in several indicators. Many investments have been made to address individual health outcomes in Mississippi with minimal return on investments. To establish measurable baselines and inform interventions, we have built a managed data lake and big data analytics platform from which to address these major challenges.

Problem. The most immediate challenge was how to manage the deluge of data. We continue to accumulate data from many different contributors from organizations across the state and collect publicly available datasets relevant to health planning.

Objective. In order to manage the lifecycle of massive amounts of data, our objective was to develop a web-based lifecycle management (LCM) application, not only for master data management, but also to manage our custom applications. Specific aims were to:

1. Design and build a Microsoft SQL database to store data files.
2. Build in security to limit access to the database.
3. Develop a front-end interface to manage data files. This web application is used to log all raw data and then track these data throughout the processing phase along with detailed revision history, and in some cases loaded into the data warehouse.

Methods. A number of technologies were used to develop the LCM application, including Visual Studio 2012, C#.Net, ASP.Net MVC5, SQL Server 2008 R2, and others. The web application was built to be scalable to accommodate all aspects of data and application management within the LCM site.

Conclusion. This custom lifecycle management application enables us to continue to build the data lake and applications, while effectively managing contributions, modifications, and custom applications. Implementing an LCM application to organize and manage all types of data is a critical first step to realizing the value and potential a big data platform has to offer.
Visual Exploration of Temporal Data in Electronic Medical Records

Josua Krause¹, Narges Razavian, PhD¹, Enrico Bertini, PhD¹, David Sontag, PhD¹

¹New York University, New York, NY

Introduction and Background

Electronic medical records (EMRs) and administrative data contain a large number of distinct events happening in the life of a patient, including diagnoses, procedures, medications, and lab test results over time. Visualizing this data would be useful for clinicians to quickly comprehend a patient’s past medical history, and for researchers to design predictive models, better understand chronic diseases, comorbidity patterns, and contributors of health care cost.

Methods

For the visual representation of a patient’s history we display every event as a dot (see Figure 1a), similar to LifeLines¹. The color of the dot represents the group of the event, i.e. a diagnosis, a performed procedure, a lab-test result, a prescribed medication, etc. The horizontal position of a dot shows the day, whereas the vertical position distinguishes between different types of events. A significant contribution is to sort the types by the first occurrence in the history of the patient. This order is helpful for using the data to form causal hypotheses. Additionally, it helps with identifying major incidents (indicated by a steep rise) and allows for the visualization of a larger number of types than in previous work. Since the space for labeling events by type is limited, we use this fact in combination with the cost of an event (indicating the importance) to select which labels to display. Additionally, a user can use an interactive “lens” to hover over events of interest to show their labels (see Figure 1b). To reduce vertical space requirements we also provide different levels of aggregation according to the CCS hierarchy for diagnosis and procedures. The software is web based using a flexible format allowing for multiple datasets as input. We developed the system using health claims from a private insurer, and also the semi-synthetic data set from CMS².

Results and Discussion

The tool proved to be helpful for analyzing the output of machine learning algorithms in predictive healthcare analysis. It provides a quick way to identify the problems of an automatically generated model by looking at patients that were classified wrong or turn out to be outliers in the cohort. Additionally, the tool enables clinicians to quickly “tell the story” of the patient using just the administrative data often found in insurance claims.

Figure 1a: Overview of the system. At the top general information about the patient is displayed. The main view shows events of the patient’s history. The view area can be zoomed into providing a more detailed view. Blue stripes indicate hospital stays. On the left the CCS hierarchy of events happening in the timeline are shown. To give an impression about the frequency and importance of a type its occurrences along the timeline are also encoded here, shown as colored lines for each occurrence. At the bottom the sum of all costs of one day are shown as histograms.

Figure 1b: A user can use a lens to get information about events she is interested in. The horizontal granularity of single events is by day and the vertical position is determined by the first occurrence of the particular event in the history of the patient.

References

Problem List Quality in Ambulatory Medicine
John C. Krauss, MD, University of Michigan, Ann Arbor, Michigan

Introduction. The diagnosis based problem list is the cornerstone of problem oriented charting, and maintenance of the problem list in structured language is a core measure of the Meaningful Use program. There is not a harmonized definition of a problem list, with ONC defining the problem list as “the current and active diagnosis, as well as the diagnosis relevant to the current care of the patient, and if there are no active diagnosis, an entry indicating that has to be made”, while The Joint Commission has a similar metric called the problem summary list which includes “any significant medical diagnosis and conditions, any significant operative and invasive procedures, and adverse or allergic drug reactions, any current medications, over-the-counter medications, and herbal preparations”. There are also different standards for the frequency of updating the problem list, with one update during the reporting period required by Meaningful Use, while TJC requires updating whenever “there is a change in diagnoses, medications, or allergies to medications, and whenever a procedure is performed.” Given the conflicting requirements by certifying entities for the definition of a problem list, the different frequencies with which the problem list needs to be updated, and the lack agreement on criteria for assigning a diagnosis to a patient, there is considerable confusion as to the use of the problem list in the EHR. 1

Methods. A chart review was conducted to evaluate how the problem list was being used after installation of a certified EHR (Epic Ambulatory). A total of 86 charts were reviewed: 71 charts at a single encounter from 9 oncologists, and 15 charts before and after an encounter from 5 primary care providers. The University of Michigan imported problem summary list items from the legacy system into Epic as structured diagnosis at the time of go-live, and these chart reviews were conducted 14-16 months after go-live. For the oncology chart reviews, a single encounter was reviewed shortly after closure of the encounter and the items on the problem list were tabulated as to inclusion of the cancer diagnosis, duplicates, signs and symptoms, and surgical history. For the primary care chart reviews, the appearance of the problem list was captured immediately before and immediately after the encounter, and changes to the problem list were noted, as well as the number of items, and the type of items included on the problem list. The accuracy of the problem list was ascertained by reviewing the entire medical record.

Results. The 14 providers wrote notes on all of the encounters; in those charts, there were 1-30 structure language diagnoses included on the problem lists. Despite the “mark as reviewed tab” being checked for 95% encounters, most of the problem lists contained errors. Those errors included 73% of the lists with duplicate diagnoses. 60% of the lists with signs and symptoms that could have been better described as a unifying disease, and 10% of the problem lists contained surgical history items. For the 15 charts reviewed before and after an encounter, 40% of the providers made a change to the problem list. For the more in depth chart reviews, all of the listed problems could be confirmed by data in the medical record, but 60% of the patients had active problems that were not added to the problem list. The problem list was not included in the soap notes of any of the providers.

Conclusions. The lack of a harmonized definition of a problem list, the difficulty in incorporating problem list maintenance into EHR workflows, and the difficulty of determining which diagnosis are significant enough to be included in the problem list all contribute to low problem list quality. A low quality problem list can have numerous adverse consequences, including inappropriate activation of clinical decision support tools, inaccurate reporting to registries, and incomplete representation of the entire patient by the medical record. Efforts to engage physicians to curate and update the problem list without adding to physician workload may improve problem list quality and provide a more accurate summary of the patient.

References
Errors with Manual Phenotype Validation: Case Study and Implications

Polina Kukhareva, MS, MPH1, Catherine Staes, MPH, PhD1, Tyler Tippetts1, Phillip Warner, MS1, David E Shields1, Heather L Mueller2, Kevin W Noonan2, Kensaku Kawamoto, MD, PhD, MHS1

1Department of Biomedical Informatics and 2University of Utah Medical Group, Univ. of Utah, Salt Lake City, Utah, USA

Abstract

Human review is generally considered the gold standard for validating computer-generated phenotypes. This work provides a cautionary case study in the context of expert chart audits conducted to validate electronic clinical quality measures for enterprise use. We found that 8.7% (176/2034) of original human review results had to be modified following initial validation. Here, we describe the most common reasons for these modifications and discuss implications for establishing a systematic electronic phenotyping validation framework.

Introduction

Electronic phenotyping entails the automatic identification of cases which satisfy specific conditions. Currently, there is no standard framework for electronic phenotyping validation. While validating quality measures for enterprise implementation at the University of Utah Health Care, our group initially developed an ad hoc validation methodology that was not sufficiently robust. We neither selected cases randomly, nor did we ensure an adequate mix of positive and negative results. To improve the quality of our validation strategy, we (1) evaluated the accuracy and reliability of our initial validation approach and (2) developed requirements for a new electronic phenotyping validation framework.

Methods

Electronic phenotyping for HEDIS quality measures was performed using a clinical decision support-based quality measurement (CDS-QM) framework [1]. Human review was performed by an individual with 16 years of chart review experience (HM). Quantitative and qualitative methods were used to analyze error types and rates for human chart review results. We employed a user story mapping technique to create requirements for a systematic validation framework.

Results

Initially, 2034 electronic numerator-denominator results were evaluated by human chart review. Among the 167 mismatches between electronic and human phenotyping, 14 (8.4%) mismatches were due to errors in human evaluation. After correcting the quality measure implementation, an additional 162 (8.7%) mismatches were identified among the 1867 results that were matched during the initial validation process. Overall, 8.4% of the humanly-reviewed results had to be changed. The reasons for errors in human evaluation differed among the HEDIS quality measure types (Figure 1).

User story mapping analysis showed that in order to provide the best results, the validation framework should enable the following: (1) choose a representative, stratified random sample of electronic phenotyping results for human review; (2) include blind validations into the sample to avoid any potential bias resulting from the validator knowing the electronic results; (3) include intermediate results to simplify validation; (4) include a mix of positive and negative cases; (5) record reviewer’s comments, credentials, and dates of review; (6) use up-to-date human review results for ongoing and automated system validation; (7) account for cases where evaluation results are expected to change in the future (e.g., when a new data source is made available), and (8) deprecate and archive old validation results.

Conclusion

Human chart audits may result in errors, especially for highly complex algorithms and the identification of rare events (e.g., rare exclusion conditions). A systematic phenotype validation framework must account for these tendencies.

References

Integrated Health Information Architecture to Facilitate State-wide and National Evidence-Based Public Health Monitoring: A Case Study based in India

Manish Kumar, M.Sc., Rohit Ramaswamy, Ph.D., and Javed Mostafa, Ph.D., University of North Carolina, Chapel Hill

Context
The fact that significant investments are made toward large-scale health programs both at national and state levels requires close scrutiny of such programs based on routine surveys and data collection operations besides increased use of data for decision making and policy formulation. The Health Information System (HIS) landscape in India, unfortunately, is highly fragmented and operates in silos. As a result, data-based health policy and program development has become tedious. Interestingly, the draft new national health policy1 of India envisions to expand universal health coverage, and strengthen health information systems for effective monitoring and assessment of policy implementation. In order to effectively meet diverse data needs of various stakeholders, accurate and timely information exchange among information systems becomes a pre-requisite. To facilitate information exchange among different levels of health system an integrated health information architecture is required. The new health policy aims to develop a health information architecture to drive evidence-based monitoring, evaluation and quality improvement efforts at all the levels of healthcare system1. In the backdrop of the health policy mandate and lack of pertinent knowledge and in-country experience, conducting a research study on health information architecture for improved policy decision making is a timely and useful exercise. For the purpose of this research study the Indian state of Bihar, one of the poorest states in the country, was selected as a case. It is a high priority state for international donors like DfID and the Bill and Melinda Gates Foundation that are implementing innovative health IT interventions like mobile Kunji and mobile academy2.

Research Methods
The objective of this research study was to analyze the current state level health information system landscape in Bihar and then suggest an integrated, but federated, Health Information Architecture (HIA) that will encourage information system innovations at the district and sub-district level while meeting the data and information needs of state level health policy makers. Besides, such an architecture will enable both vertical and horizontal scalability. Two principal research techniques namely secondary literature review and primary data collection were used in this study. The literature search was conducted in two scientific databases namely PubMed and WebofScience. The primary data collection involved eight key informant interviews in Bihar.

Results and Conclusions
The research findings highlight the fragmented HIS landscape and discuss the a) healthcare provider landscape, type of clinical data collected and gaps in clinical data, b) customer’s socio-economic and demographic profile, access to health care services, type of clinical data and interaction with the providers, c) key drivers and motivators for the policy makers and politicians, data collection and analysis, use of data and related decisions, health information technology infrastructure, human resources, and regulations. Based on these findings, a federated HIA model is proposed that can serve information needs of both healthcare providers and public health leaders. Drawing from the global experiences, especially the health information exchange experience in the United States, the final section of the paper describes the factors that are most likely to influence implementation and sustainability of the federated HIA model in a low resource setting.

References
Archetype Based Nationwide Electronic Health Record Development in Japan

Naoto Kume, Ph.D1, Shinji Kobayashi, M.D.1, 
Kenji Araki, M.D.,Ph.D2, Satoshi Inoue3, Hiroyuki Yoshihara, M.D.,Ph.D2 
1EHR Research Unit, Kyoto University, Kyoto, Japan; 
2Miyazaki University Hospital, Miyazaki, Japan; 
3Japan Medical Network Association, Tokyo, Japan

Introduction

Successful installation of electronic health record (EHR) relies on government initiative. However, public sector does not have affordance to develop and sustain electronic health record as a lifelong record repository in Japan. Billions of subsidies are invested to establish domain specific clinical data repository in the last two decades. Even though, EHRs are not promoted except for a few regions. New government policy, control of increasing medical expense and international competitiveness of medical research, leads the direction that secondary use of medical data should be promoted. Now is an occasion that an update of Personal Information Protection Law and a new national identification number law are approved in the Japanese National Diet. In these circumstances, the authors develop an EHR system and management organizations, by renovation of conventional EHR system under operation for the last decade.

Purpose

The authors run EHRs in Kyoto and Miyazaki region last 15 years using an open source iDolphin, which is implemented by nonprofit organization Japan Medical Network Association (JMNA). This study aims to update iDolphin from scratch based on ISO 13606 Archetype standard, instead of iDolphin original standard medical markup language (MML), which is defined by Medical eXtensible Markup Language Consortium. Also, establishment of two types of management companies for data collection and secondary use are proposed.

Methods

The authors propose a medical information circulation model by establishing EHR cloud system and secondary use systems. Because unification of several standards requires continuous struggles, the authors allow multiple standards for data exchange. Internal database of the proposed system, which is established in a cloud data center, is designed to store the data by ISO 13606 Archetype standard. For instance, EHR systems are implemented by OpenEHR, which is an extension of ISO 13606 Archetypes. In the same way, to optimize Archetypes for Japanese localization, MML is decomposed and reassembled by OpenEHR Archetypes. For example, Archetypes originated from MML are defined for vital sign, prescription and so on. The proposed EHR cloud system, so called global Dolphin, and secondary use database are separately installed and managed by different organization because of compliance with privacy policy. Secondary use contains for clinical support use, patient-informed use, and research use. The secondary use sector gains profit from users so that the profit follows EHR cloud system sustainability. Primary target of data collection is text data; laboratory test, prescription, insurance claim, and some reports such as discharge summary.

Results

This project is funded by the Japanese Cabinet next generation medical ICT council. Data source facilities, including university hospital, public hospital, health examination center, laboratory test center and dispensing pharmacy, have different policy to connect EHR. Some facilities fully agree with the concept. Another require direct profit before connecting EHR, even though it cannot be expected until secondary use database is completely operational. The others emphasize security risk of unexpected information disclosure and do not participate.

Conclusion

Nationwide EHR is desired from patients, clinical use, and research especially for drug manufacture. Japanese government promotes to establish use cases of collected medical data. The authors are under development of nationwide EHR based on an international standard Archetypes. As medical and healthcare organizations' policies have to be paid serious attention, the proposed model is proceeded proof-of-concept with early collaborative organizations.
Integrating Usability Engineering into Undergraduate and Graduate Health Informatics Curricula

Andre W. Kushniruk, MSc, PhD, FACMI, Elizabeth M. Borycki, RN, PhD

1University of Victoria, Victoria, British Columbia, Canada

Abstract

Usability is a major issue in the successful design and implementation of health information systems. However, practical training in usability engineering methods has been limited in health, medical and nursing informatics programs. At the School of Health Information Science at the University of Victoria, an approach known as rapid usability engineering has been applied in a wide range of health informatics projects and teaching contexts. The poster describes how the approach has been applied within our curricula and in work post-graduation.

Introduction

The success of healthcare information technology (HIT) depends on the usability of that technology. However, there are many reports of systems failing due to poor usability. It is clear that there are gaps in the education of HIT professionals about usability. At the University of Victoria’s School of Health Information Science we have been working to address this knowledge gap by introducing practical approaches to usability engineering within our undergraduate and graduate curricula. The approach taken has been based on development of methods that we have designed – namely an approach we have termed “low-cost rapid usability engineering” 1. The approach takes a pragmatic view of usability engineering in HIT that can be conducted in-situ, at low cost and rapidly in real clinical settings. In this poster we outline our approach to bringing usability engineering into health informatics curricula.

Methods

During a major curricula review a qualitative analysis was conducted to determine where usability engineering could be incorporated in our curricula. Based on interviews and surveys with key stakeholders it was determined that a third year undergraduate course entitled “Human Aspects of Healthcare Information Systems” would be an ideal insertion point to embed 4 month-long small group projects that required students to apply the usability methods. At the graduate level the methods were embedded in a course on human factors for HIT. Information about the positions our students took during co-operative educational experiences and upon graduation was also analyzed.

Results

465 undergraduate and 143 graduate students in our health informatics programs received training in the application of rapid low-cost usability engineering methods since 2008. The main pedagogical approach to their training has been through project work requiring application of low-cost rapid usability engineering on real world projects. This has lead to application of their skills in a variety of co-operative educational experiences (where our students have introduced the methods to their employers) and the new application of the methods in key organizations where students have taken up key positions upon graduation (including in the government, industry and hospital sectors). Students perceived the course as being useful and this increased over time. Interviews with graduates indicated over 50% applied their learning in a variety of roles (e.g. as system analysts, managers), while a growing number were working as human factors specialists in healthcare (e.g. working overseeing usability testing of EHR deployments).

Conclusion and Discussion

In our work we have attempted to integrate competencies related to usability engineering into undergraduate as well as graduate health informatics programs. The approach described has worked to create a cadre of health informatics professionals who have not only gained understanding of the importance of usability, but who have hands-on practical experience in applying usability engineering methods and who have worked in transferring that knowledge and skill in a range of sectors. In addition, the topics are evolving to include usability of mobile and other devices.

References

Information Acquisition Preferences in the Intensive Care Unit

Kathryn G. Kuttler, PhD1,2,3, Jorie Butler, PhD2,4, Eliotte L. Hirshberg, MD1,2,5,6, Ramona O. Hopkins, PhD1,2,7, Emily Wilson, MStat1,2, James Orme, MD1,2,5, Samuel M. Brown, MD MS1,2,5

1Pulmonary and Critical Care Medicine, Intermountain Medical Center, 5121 S. Cottonwood St, Murray, UT
2Center for Humanizing Critical Care, Intermountain Healthcare, 5121 S. Cottonwood St, Murray, UT
3Homer Warner Center for Informatics Research, Intermountain Healthcare, 5171 S. Cottonwood St, Murray, UT
4VHA GRECC Salt Lake City, Utah; University of Utah School of Medicine, 30 N 1900 E, Salt Lake City, UT
5Pulmonary and Critical Care Medicine, University of Utah School of Medicine, 26 N 1900 East, Salt Lake City UT
6Pediatric Critical Care, University of Utah, 26 North 1900 East, Salt Lake City, UT
7Psychology Department & Neuroscience Center, Brigham Young University, 1022 SWKT, Provo, UT

Background

Intensive Care Units (ICU) are stressful environments for patients and their family members. Potentially life changing information inundates patients and their family members, who often suffer from Post Traumatic Stress Disorder (PTSD) after discharge. Although the reasons for distress in the ICU are likely multifactorial, in several studies family members of ICU patients reported information needs and decision related stress as important contributing factors (e.g., McKiernan & McCarthy, 2010; Azoulay et al., 2005; Botti, Orfali, & Iyengar, 2009). The modality of information delivery could be influenced by the personalities and communication preferences of the family members of ICU patients. People who want to use computer tablets may have different personalities and communication preferences.

Methods

Participants from the Osher Institute for Lifelong Learning in Salt Lake City and college students from the University of Utah were asked about their preference for using a tablet through the administration of the following: 1) a modified version of the Degner Control Preferences Scale (CPS) (Degner, Sloan, & Venkatesh, 1997) 2) The abbreviated Big Five Inventory (BFI-10). The CPS was used to measure preference for engaging in shared decision making. Participants indicated their preference from most preferred to least preferred answering, “how you would prefer to make decisions with your doctor” with the mid-point of the scale representing an equally shared option. The BFI-10 was used to assess neuroticism (mean of responses to the following statements: “I see myself as someone who is relaxed, handles stress well.” [reverse scored] and “I see myself as someone who gets nervous easily.”). In other samples, the reliability and validity were shown to be good (Rammstedt & John, 2007).

Results

Participants’ (N=359) data were analyzed. One outcome measured correlations between “tablet use” (higher numbers mean more willing to use a tablet to obtain information) and the following covariates: age, sex, preference to utilize the Internet when seeking medical information, the BFI personality variable—neuroticism, and a preference for shared decision making using the CPS. Using the Internet to obtain medical information is positively correlated with “tablet use” (r = 0.11, p < 0.05); neuroticism is negatively correlated with “tablet use” (r = -0.11, p < 0.05); and CPS is negatively correlated with “tablet use” (r = -0.12, p < 0.05). This implies that a participant with relatively high neuroticism and higher scores (signifying a more deferential preference for participation in shared decision making) on the CPS is less willing to use a tablet. A multivariate linear regression analysis showed a preference for more deferential decision making (β = -0.11, p < 0.05) and researching health information using the internet (β = 0.30, p < 0.05) outperformed the other variables in predicting preference for tablet use, indicating that information preferences are more important than neuroticism, age, or sex (r² (adj) = 0.03; F (df, df) = 3.343(5,342); p < 0.01).

Conclusion

Study participants report a willingness to use a computer tablet. Tablet users have a preference for more autonomous shared decision making, and researching health information using the internet. These findings would indicate that an ICU transitioning towards a patient-centered model should have tablets available for patients and families to allow for greater access to information and the internet. These information tools could assist patients and families to explore care options and make treatment decisions with their clinical team.
Evaluation of Simulated Computerized Provider Order Entry Rules Toward Evidence Based Blood Utilization

Jaie Lavoie, PharmD¹, William V Miller, MD², Min Kim, PhD¹,³
¹Department of Health Management and Informatics, University of Missouri, Columbia, MO; ²Department of Pathology, University of Missouri School of Medicine, Columbia MO; ³Informatics Institute, University of Missouri, Columbia, MO

Abstract
We evaluated how additional alert override steps affect evidence based and non-evidence based transfusion ordering behaviors. Nine inpatient physicians participated in a newly created simulated ordering environment, where five providers received a typical scenario for overutilization and four providers received a scenario where a rule that only considers hemoglobin would discourage an evidence based transfusion. The new order rules prevented 100% of inappropriate orders while allowing 75% of appropriate orders to override an incorrect alert.

Introduction
Studies suggest that as much as two thirds of red cell orders are non-compliant with evidence based recommendations. CPOE rules based on patient hemoglobin levels have been a proposed solution to this problem. However, it is unclear how difficult to override a rule should be in order to be effective at preventing orders with no evidenced based indication and what the potential impacts of these rules are on necessary transfusion ordering. The purpose of this pilot study is to evaluate the effect of additional alert override steps on evidence based and non-evidence based transfusion ordering behaviors.

Methods
A web based application was developed to simulate order entry. The application presented two medical case vignettes and then simulated an interactive order entry process with a triggered alert. The alert rules were solely based off the most recent hemoglobin value with a cut point of 8 mg/dl to determine if an order was in compliance with evidence based practice. Attempting to override the alert required providers to enter free text justifications with at least 150 characters to continue. The two clinical vignettes designed were: (1) an overutilization vignette having characteristics of a case that in practice would commonly receive a transfusion, however not have an evidence based indication, and (2) a non-hemoglobin indication vignette with an evidence based indication for a transfusion that is not likely to be detected by automated rule based systems. In this manner the former would be more likely to prompt a non-evidence based order and desired alert while the latter would be more likely to elicit an evidenced based order but an undesirable alert. Simulation data, including orders, number of alerts generated, attempts, justifications, and success at overriding an alert were recorded into a relational database.

Results
Nine subjects were recruited from a pool of eighty seven inpatient physicians who were eligible to order a blood transfusion at the university health system. Five providers were randomized to the overutilization vignette and four to the non-hemoglobin indication vignette. Three out of five providers (60%) in the overutilization vignette chose to order a blood transfusion and were given an alert that the order was not recommended based on the hemoglobin level. One provider canceled the order immediately while the other two providers attempted to override the alert. Upon being presented with the free text justification box that could not be left blank, however, the two providers who initially attempted to override the alert withdrew their orders. On the other hand, all four providers who received the non-hemoglobin indication vignette, placed an order for a transfusion. These four providers were given an alert that the hemoglobin value did not indicate a transfusion as the indication was not based on hemoglobin but instead a free text finding in the note not considered by the rule. While three of four providers (75%) successfully wrote an appropriate justification and overrode the alert, one chose to cancel the order after unsuccessfully attempting to submit a justification below the character length requirement.

Conclusions
While data collection is still in progress, preliminary results suggest that rule based alerts on patient hemoglobin can be highly effective in preventing unnecessary blood transfusion orders especially when they require an additional evidence-based review process. For example, two out of three non-evidence based orders attempted to override, which had the process been as simple as a single click would have allowed these orders to continue. While preventing unnecessary orders is critical, care must be taken to not prevent orders indicated by non-discrete clinical findings or where the hemoglobin is rapidly trending down / outdated. Acknowledging the limitation of the current alerts used in CPOE systems, a balance should be made between the ability to deter unnecessary orders and the ease of overriding erroneous alerts. This study was conducted in a non-clinical setting in a single center with a small sample which warrants large study in a real clinical setting. For example, higher levels of stress/alert fatigue may alter provider behavior in a real clinical setting, which may not be observed in the study. Despite these limitations, the preliminary results are encouraging and merit continued study towards the more effective means of leveraging CPOE systems to help optimize effective blood management in the health system.
Semantic and Interactive Timeline for Patient Data Visualization

T. Ledieu1,2, P. Van Hille1,2,3, G. Bouzillé,MD1,2,3, E. Renault1,2, M. Cuggia,MD,PhD1,2,3
1INSERM, U1099, Rennes, F-35000, France; 2Université de Rennes 1, LTSI, Rennes, F-35000, France; 3CHU Rennes, Centre d’Investigation Clinique, Unité Fouilles de Données, Rennes, F-35000, France

Introduction
Considering the multiplication of sources of health care information, providing to physicians effective and integrative tools to easily visualize the medical history of the patient is crucial. A timeline is a chronological representation of a list of events which may use different time and data scales. In this work, we introduce new advanced features into a timeline framework for patient data visualization, using interactive GUI, semantic data aggregation and time navigation functionalities. The purpose of this visualization tool is to offer a generic and synthetic display of the medical data which could be adapted according to a specific medical problem or the physician specialty.

Methods
The prototype was developed from an existing timeline component based on D3 Framework1. This prototype was designed as a weakly coupled component in order to be integrated to different data repositories such as EHR or Clinical Data Warehouses. The patient dataset used for the development includes free text, structured and coded data with different international and French terminologies (e.g. ICD-10 for diagnosis, ATC for drugs, CCAM for procedure codes), and numerical data (results lab). A semantic aggregation method uses hierarchical properties of terminologies. Domain and visualization ontologies were developed and combined with a set of rules to dynamically filter the patient dataset and generate adapted views according to the physician specialty into the timeline.

Results
The user can navigate in time using a temporal window and choose the time scale in those synchronized timelines. The user can search a medical concept in a free-text zone, and select it to display the data into the timeline. This data is shown by colored dots for punctual events or colored rectangle for a period. The user can aggregate/disintegrate the data element under the different classification systems. He is also able to read textual documents and search terms in full text. A specific displayed have been developed for one medical field: rheumatology. The rheumatology timeline shows the evolution of patient’s joints status on a diagram. In order to make a clearer understanding display for a clinician, we have integrated an icon-based language (VCM language²) mapped on the ICD-10 classification to display diagnosis in the timeline.

Conclusion
Timeline representation helps to integrate multi domains patient information over time. Our prototype is used by both physicians and data scientists to better analyze health care events in several dimensions. This work is in progress: we plan to represent textual data through temporal dynamic semantic clouds. Conceptual navigation feature will be extended to use other kind of semantic relationships (e.g. part_of, has_finding). Aggregation visualization method will take into account temporal dimension. A search engine will be developed for enable the composition of a customized timeline by users (e.g. timeline for a specific medical problem or several timelines for comparison of patient care trajectories).

References
The Master Data Element Visualization: A Consolidated View of the EHR Data at Intermountain Healthcare

Jaehoon Lee, PhD1, Farrant H. Sakaguchi, MD1, Justin Mundt, MS1, Bart B. Dodds1
Gary D. Bishop1, Nicole K. Hobbs, PhD1, Katherina C. Holzhauser1, Stanley M. Huff, MD12
1Intermountain Healthcare, Murray, UT; 2University of Utah, Salt Lake City, UT

Abstract

The master data element visualization provides navigation and interactive visualizations of the electronic health record (EHR) data at Intermountain Healthcare. The key features are 1) a sunburst tree visualization to represent classes of hierarchies, 2) data visualization for volume and distribution of the EHR data, and 3) a search function for concepts and codes. The tool was implemented internally for clinical researchers and demonstrated its usability.

Background

Intermountain Healthcare is a non-profit healthcare organization that has collected 20 years of clinical data through its EHR, and visualization has been used in clinical applications to provide an intuitive understanding of the data. However we have several challenges to 1) handle the size, complexity, and contextual understanding of huge EHR data, 2) develop normalize scheme for aggregated data, and 3) present a large set of information in a single screen.

Method

We developed the Master Data Element visualization: a web-based data navigation and visualization tool based on d3.js1. The tool deals with essential data elements of the EHR such as patient demographics, encounters, vitals and labs, diagnoses, procedures, and medications. To provide visualization of large population in a timely manner, we developed an aggregated data model and cached the data of 4,676,644 patients from 1994 to 2015. The key features of the tool are 1) a sunburst tree visualization to provide a high-level overview and drill-down of Intermountain’s data hierarchies (Figure 1 Left), 2) a search function for a concept or a code in the hierarchies (Figure 1 Left), and 3) 16 data visualizations using four types of charts (line chart, bar chart, box plot, and histogram) to present volume of codes, number of patients, and statistical distribution of numeric values (Figure 1 Right).

Figure 1 Left) By clicking a slice of the sunburst tree users can drill down the hierarchies. Users can search a concept or a code on the right panel and each search result provides a hyperlink (red) to data visualization of the code. Right) At a selected category (e.g. vital sign) of the hierarchies, the right panel shows available data visualization; each visualization provides drill down at code level (e.g. histogram of respiratory rate of patients by sex, age group).

The tool was implemented within Intermountain’s firewall to be used for internal clinical research and demonstrated its usability, computational efficiency, and feasibility to address the problems stated above to some extent. However the tool currently covers only small part of Intermountain’s EHR and it taxonomy is limited to present complex hierarchies of the EHR data. Our future work is to adopt knowledge bases and standard terminologies from Intermountain’s healthcare data dictionary to enrich the hierarchies and the search functionalities.

References

Applying an instant messaging system at the hospital to support TRM

Ying-Li Lee¹,², Tsai-Feng Chien¹, Hsiu-Chin Chen¹
¹Department of Nursing, Chi Mei Medical Center, Tainan, Taiwan; ²Institute of Biomedical Informatics, National Yang Ming University, Taipei, Taiwan

Abstract

Effective communication among the healthcare team is needed to facilitate team resource management (TRM). We aimed to develop an instant messaging system, Link to Team (L2T), embedded in the original hospital information system to improve the communication and coordination among all members of the healthcare team. Based on the idea of patient-centered and workflow-driven, we used the rapid application development model (RAD) to develop an easy-to-use and user satisfied instant message system.

Introduction

Nurses spent 19-24% of their time in professional communication.¹ They often connect and communicate with different members of the healthcare team by telephone. However, the telephone communication might interrupt the current work of the receivers. The misunderstanding of messages could happen during message transcribing or remembering incorrectly. We developed an instant messaging system which was nicknamed Link to Team (L2T) system. The purpose of our project was to support the healthcare team resource management and improve the communication and coordination among all members of the healthcare team.

Method

We conducted the focus group interviews, the healthcare teams’ workflow surveys and the workflows simplifications before designing the system. According the text message proposed by users themselves, we set up various message templates in each communication scenario. The rapid application development model (RAD) was applied to software development. After several iterative modifications, the system was implemented officially. After 3 months of implementation, we conducted interviews, a system operation testing and a satisfaction survey for the users to evaluate the preliminary outcome. After 6 months of implementation, we analyzed the usage of the system.

Result

In preliminary outcome evaluation, we interviewed 15 nurses. The respondents indicated that they have replaced most telephone communication by the L2T system. It’s very convenient to transmit a message to a number of recipients simultaneously. It saved lots of time in team members connecting. Besides, the content of the text messages was more correct and clear. Users can also query history information in the system to avoid forgetting messages. We also randomly sampled 29 nurses and conducted a questionnaire survey for primary users’ experience. Most users were satisfied with the color scheme and the design of the interface, easy to learn and easy to use the system. 86.2% users were familiar with the system function. After 6 months of implementation, we analyzed the history records of the system from January 1 to January 31, 2015. The total number of records was 20,781. 51.21% messages were related to notifying another nursing station about patient transfer information. 31.32% messages were communicated between the ward and operating room about the patient surgery preparation. 13.38% were used for communication between colleagues of the same nursing station. The outpatient clinic nursing staffs have the highest utilization rate in requesting support. 68.42% of messages were received and processed within three minutes.

Conclusion

This project was satisfied the users’ expectation in rapidly, easily and correctly operative. Meanwhile, it also improved the timeliness and coordination between the healthcare team. The L2T system was designed as a communication tool to meet the workflows of various departments. Next, we would like to integrate the workflows of those departments that have not been included in this project. It is expected that wider adoption of the current L2T system will help to improve clinical and administrative work processes in the future.

References


1548
Quantifying the Complexity of Discharge Planning in the Inpatient Cardiology Unit

Young Ji Lee, PhD, RN¹, Corrine Benacka, RN, CCRN², Matthew B. Carson, PhD¹, Gayle Kricke, MSW¹; Preeti Kansal, MD², Nicholas Soulakis, PhD¹

¹Northwestern University, Chicago, IL; ²Northwestern Memorial Hospital, Chicago, IL

Introduction: Well-organized, multidisciplinary discharge planning bridges the gap between home and hospital to reduce readmission rate²-³. Patients should have the option to be active participants in the process; however, they are often confused whom they can ask for help due to the complexity of the process¹. A study describing discharge planning found patients typically meet 3 to 7 types of healthcare provider; however, to our knowledge, no study has quantified the actual number of providers involved in discharge planning². Electronic Health Records (EHR) has become an essential research resource for exploring the structure of healthcare providers³. Thus, the goal of this study is to understand the complexity of discharge planning in an inpatient cardiology unit through EHR by quantifying the total number of providers and the number of new providers per day involved in discharge planning activities.

Methods: Using the Northwestern Medicine Enterprise Data Warehouse (NM EDW), we identified records of patients with heart failure (HF) who were admitted to Northwestern Memorial Hospital’s cardiology unit and discharged between 12/1/14 and 12/31/14. All HF-associated providers and discharge-planning activities were also extracted. “Discharge planning activities” were determined and validated by a nurse care coordinator and the director of the cardiology unit. We defined “Providers” as healthcare providers and non-clinic personnel who performed discharge-planning activities. “New providers” were defined as a provider not previously seen at the point of the activity. Terminologies followed the Cerner system. The analysis was performed using custom scripts written in the R program.

Results: We identified 42 patients, 99 providers and 839 individual discharge-planning activities. Sixteen types of providers; 19% of providers were ‘Patient care staff nurses’. Thirty discharge-planning activities were identified within the timeline. The range of patients’ length of stay was between 1 to 38 days (median=6). All 42 patients started their discharge planning at the first day of their admission, and they met 7 providers for their discharge planning on average. We found out that patients met most discharge-related providers near the end of their hospital stay (Figure 1). Interactions between patients and providers are visualized in Figure 2.

Conclusion: We found that more types of providers are involved in discharge planning than stated in previous literature. Also, the study showed that the number of providers accelerates towards the end of an inpatient stay. Further study with a larger sample is needed to understand how complexity of care affects health outcomes.

Reference

Design of a Knowledge Exchange for Community Health Workers

Harold P. Lehmann, MD PhD, M. Christopher Gibbons, MD MPH, Richard Singerman PhD, Steven King, Joe Warren MS, John May MS MBA, Johns Hopkins School of Medicine, Baltimore, MD; Johns Hopkins School of Public Health, Baltimore, MD
TrustNetMD, Inc.

Community Health Workers (CHWs), a mainstay of healthcare in low- and middle-income countries\(^1\) are poised to become an increasingly important part of the US healthcare landscape\(^2,3\): They engage patient in the community, they are less costly than licensed healthcare providers, and they are effective.\(^4\) Their information needs have not been supported formally so few information resources have been provided to them that represent evidence related to those needs. Those that have focused on this issues have reported positive results using rich media to support the workflow of our CHWs.\(^5,7\)

We have created such a resource (CHWresources.org) based on the following principles: (1) CHW are experts in being CHWs; (2) CHWs have specific tasks that fall into specific classes; (3) Resources for CHWs should be tailored to those tasks

To this end, we conducted focus groups (IRB # NA_00085655) of 36 CHWs in Baltimore, Connecticut, and Brooklyn, NY to create a taxonomy of tasks falling into 4 classes: Assess client needs; Connect clients to community resources; Overcome family barriers; Attend to personal professional needs. Resources fall into categories as well: Evidence-based published articles and guidelines; other authoritative information (published texts, web sites, videos); community-based resources; and knowledge exchanged among CHWs. These classes have been validated by several CHW groups.

A knowledge exchange (Kx), also called Web 2.0 with a learning social network, enables CHWs to ask questions of each other; to rate resources or to vote questions and answers “up” or down. These features enable crowdsourcing and the emergence of expertise from among the group:

- The four task classes are denoted by color-consistent buttons and tabs.
- Clicking on either brings the user to a search page enabling typing, voicing, or selecting tags, coded according to the task classes.
- The search-results page enables rating, voting, commenting, questioning, answering, and sharing of new resources.
- The resources database comprises hundreds of authoritative and community resources.

The Web 2.0 site can be contrasted with a Web 1.0 version, which offers the information resources, but not the interaction. While it would seem obvious that a Kx is desirable, it is not clear whether the added task burden and complexity is worth the information gained. For this purpose, we are conducting a cluster-randomized trial of the Web 1.0 vs Web 2.0 versions in cities across the country (IRB #IRB00049686).

References:

Acknowledgements: Funded by AHRQ R24 HS022073. Thanks to the East Baltimore Medical Center, Health Care Access Maryland, Lutheran Healthcare, Southwestern AHEC for hosting the focus groups. Thanks to the Welch Library, especially Claire Twose, and Andrew Boxill, Angeline Aringo, Christian Minter, Neha Goel, Senyo Norbey, in collecting authoritative resources. Thanks to Healthify, Inc., for collecting community resources. Thanks to our Internal Methodology Advisory Board and to our External CHW Advisory Board for guidance.
Theme: Interactive Systems Track: Applications of Informatics
Decision Factors Influencing the Selection of a “Hand-off” Model Versus a “Hold-on” Model for Telehealth Service Lines

Cynthia LeRouge, PhD¹, Suzanne Wood, PhD¹, Ryan Sterling, MPH/MSW¹
¹University of Washington, Seattle, WA

Abstract

The need for robust and effective telehealth programs (often composed of multiple service lines) has never been greater, leaving healthcare organizations clamoring to develop delivery systems that result in effective workflow and desired outcomes. Examined are two distinct management approaches to telehealth service lines service lines - “hand-off” versus “hold-on” models, Results specify factors underlying model selection and the role of telehealth departments in facilitating effective workflows to support successful implementation and sustainable operations.

Introduction

Given the increasing number of telehealth programs, organizational structures (e.g., departments) with dedicated leaders to manage multiple service lines are increasingly common. Selecting telemedicine management models to drive appropriate workflow can prove challenging for telehealth programs due to wide variation in care settings, provider types, and service lines across which programs are implemented. Disruptions to clinical workflow can impede the success of health information technologies. Unfortunately, research that describes, analyzes, and discusses the impact of alternative management models on the implementation, adoption, and on-going operations of telehealth service lines is scarce. This study explores two distinct management models deployed for separate service lines: a decentralized “hand-off” model and a centralized “hold-on” model, respectively (see Table 1). Study objectives are to identify factors for selecting each mode, and to provide insight into the role of the telemedicine department in facilitating implementation and operational workflow under each model.

Methods

A mixed methods approach was used to perform a comparative longitudinal case study of two telemedicine programs, TeleHepatology and TeleStroke, operated at a not-for-profit health network under a single telehealth manager. Data sources included stakeholder surveys, archived program documents, clinical program results, and transcripts from program management meetings. These rich data sources enabled researchers to triangulate findings and enhance research validity and reliability. All interviews, meetings, and discussions during site visits were audiotaped, transcribed, and reviewed for errors. Qualitative analyses were conducted with NVivo © and Dedoose © using guiding principles from Lee and Baskerville (2003) to transform data into themes and develop insights1.

Results

Preliminary results identify workflow factors specific to each individual management approach as well as common factors that should be prioritized across both models. Factors unique to the “hand-off” approach stress stakeholder engagement, mentorship of service unit and telehealth managers, as well as information sharing and education. Factors unique to the “hold-on” approach stress appropriate pacing of program growth (including human resources), development of clinical champions, and assembling expertise from within when needed. Common success factors in both models stress the development of standardized, evidence-based best practices and progress assessment.

Conclusion

Results point to certain key decision factors that can guide healthcare organizations in choosing telehealth management models to support implementation and operational workflow. Specifically, variations in the capacity of service lines, technical support needs, and value proposition alignment are highly influential to decision-making.

References


Table 1. Review of key model components

<table>
<thead>
<tr>
<th>Key Model Component</th>
<th>“Hand-off” (TeleHepatology)</th>
<th>“Hold-on” (TeleStroke)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact</td>
<td>Provider-to-Provider</td>
<td>Provider-to-Patient</td>
</tr>
<tr>
<td>Care Delivery</td>
<td>Regularly scheduled sessions</td>
<td>As needed in emergency department</td>
</tr>
<tr>
<td>Telehealth Department Role</td>
<td>Supporting services enabling a service line department to lead implementation and operational workflow</td>
<td>Centralized model where telemedicine department leads implementation and operational workflow</td>
</tr>
</tbody>
</table>
Drug-Disease Associations in Guidelines, Drug Labels, and Practice

Tiffany I. Leung, MD, MPH\textsuperscript{1}, Michel Dumontier, PhD\textsuperscript{2}

\textsuperscript{1}Division of General Medical Disciplines, Stanford University, Palo Alto, CA; \textsuperscript{2}Stanford Center for Biomedical Informatics Research, Stanford University, Palo Alto, CA

Introduction
Clinical practice guidelines (CPGs) recommend pharmacologic treatments for certain clinical conditions based on systematic reviews of medical literature, and drugs' structured product labels (SPLs) provide drug profiles, including treatment indications, approved by the FDA after review of clinical trial evidence. Both resources are intended to promote evidence-based medical practices. Practice patterns should also match guideline-recommended and FDA-approved prescribing. For example, denosumab, a monoclonal antibody that inhibits bone resorption, has FDA-approved indications for postmenopausal women with osteoporosis and for increasing bone mass in men with osteoporosis. Conversely, osteoporosis guidelines typically recommend denosumab as a pharmacologic treatment option for the condition, but other disease guidelines do not recommend denosumab as a treatment alternative. In this instance, the CPG-recommended pharmacologic therapy and the SPL treatment indication match, and electronic health record (EHR) data should reflect a prescription for denosumab only for a patient who has diagnosed osteoporosis. In this study, we use text mining methods to determine how well drug product labels' treatment indications and prescribing practices for 15 chronic conditions match CPG-recommended pharmacologic treatment.

Methods
We focused on 15 common chronic conditions among Medicare beneficiaries, defined by 448 ICD-9 codes. In order to assess how well-matched drug-disease associations were across CPGs, SPLs, and practice data, we built drug-disease associations from SPLs using a computer-readable side effect resource (SIDER), outpatient prescribing data for patients with any of the ICD-9 codes corresponding to the 15 chronic conditions on their problem lists using EHR data from the Stanford Translational Research Integrated Data Environment (STRIDE), and CPG text from 494 relevant guidelines from the National Guideline Clearinghouse. For CPGs, we applied a text-mining algorithm similar to that used in a previous study to extract chronic disease mentions;\textsuperscript{(1)} however, we used this algorithm to identify drug mentions in CPG text by exactly matching a list of drug names. We used DrugBank as an encyclopedia to identify a list of more than 11,609 drug names, their synonyms, and drug categories, which we subsequently enhanced with 420 additional terms using four ontologies (SNOMEDCT, NCIT, NDF-RT, CHEBI) from Bioportal at Stanford University’s National Center for Biomedical Ontology, to create a list of 12,030 drug names. We determined the frequency of matches and mismatches of drug-disease pairs among CPGs, SPLs, and practice data.

Results
Our results yielded 1,999 drug-disease associations from 494 relevant CPGs, and 533 drug-disease associations from SPLs. Of these, there were 240 drug-disease association matches between the two data sources, representing only 12% of all associations in CPGs and 45% of all associations in SPLs. The frequency of drug-disease associations in CPGs, SPLs, or both varies depending on which chronic disease guidelines are of interest. Initial evaluation of this method for 10 heart failure guidelines for drug names alone demonstrated precision 0.98, recall 0.77, and F-measure 0.86; for drug names plus drug class names, precision was 1.00, recall 0.19, and F-measure 0.32. In EHR data, 13,689 drug-disease associations were identified and there were 1,193 matches out of the 1,999 CPG drug-disease associations. This represents 59.7% of all associations in CPGs and 8.7% of all associations in EHR data.

Conclusion
Our preliminary results suggest that there is a mismatch between guideline-recommended pharmacologic therapies and drug product label indications, suggesting potentially conflicting prescribing guidance from CPGs and SPLs. Further investigation of the identified mismatches may guide harmonization of clinical knowledge sources. The higher number of matches in EHR data compared to CPGs suggests that prescribing practices match guideline recommendations better. Further improvement of the text mining methods, particularly leveraging the taxonomic organization of drug ontologies to improve recall, may address limitations of our approach to drug-disease associations identified in CPGs and SPLs.

References
Developing an Electronic Survey to Capture Current State of Acute Care Patient Portals to Inform Best Practices and Future Directions

Waiyin Leung, MS1, Sarah Collins, RN, PhD1,2,3* and Anuj K Dalal, MD1,2,3*

1Brigham and Women’s Hospital, Boston, MA, 2Partners HealthCare System, Boston, MA, 3Harvard Medical School, Boston, MA *Co-PI

Abstract: To date, patient portal research has been in the ambulatory settings. There are limited data on the use of patient portals in acute care settings. Using an iterative process, we developed an electronic survey that aims to capture critical data on the current state of acute care patient portals at four academic medical centers in order to identify new directions for development and investigation. We describe the survey development process and the resulting survey.

Introduction: Meaningful use requirements for patient engagement and increasing patient expectations for online access to health data call for the need for expanding use of patient portals. Though evidence-based practices for patient portals are well in ambulatory settings, these are lacking in the acute care setting. Consequently, many patients and their caregivers generally do not access personalized health information via patient portals in hospital settings, although use could lead to higher patient satisfaction and better health outcomes.1 Engaging hospitalized patients through the use of health information technology in the acute care setting is an area of active innovation and development. More research regarding how institutions design, develop, and implement acute care patient portals is necessary to inform best practices. A thorough understanding of currently available infrastructure across institutions is required. Specifically, investigation should address portal content and security, policies regarding data access and availability, and management. The purpose of this study is to describe the process by which we developed an electronic survey to 1) identify key features of an acute care patient portal that are common across sites, 2) recognize common operational and implementation challenges, and 3) define best practices for portal design, configuration, and use for future investigation and innovation.

Methods: We developed a survey using REDCAP2 (Figure 1), an electronic data capture tool hosted at Partners Healthcare Systems. Questions were modified based on a survey instrument that assessed policies and current state of ambulatory patient portal.3 We included questions regarding both inpatient and ambulatory patient portals given the growing interest in developing enterprise-wide patient portals that provide a unified patient-centered experience for patients and caregivers across all settings. We used a participatory process with input from collaborators at four academic medical centers to iteratively refine the survey questions.

Results: Because study investigators were concerned that the survey did not consider institutions that do not have an existing acute care patient portal, we included a question to assess whether an individual institution had an existing acute care patient portal; if not, they were asked if they were planning to acquire, develop, and implement one in the future, and if so, they would provide insights on their plan. The survey used branching logic and multiple formats including yes/no questions and free text boxes to capture quantitative as well as qualitative data, respectively. The final survey asked questions regarding patient portal leadership, access for patients and healthcare proxies, usage statistics, key features and clinical data available for access by patients, specific policies in place, challenges encountered/anticipated during development and implementation, and visions and goals for improvement or development.

Conclusion: We developed an electronic survey that captures a wide range of details about a healthcare institution’s strategy for future acute care patient portal use in order to identify common practices across institutions. The survey addresses a critical need for establishing best practices and future direction for acute care patient portal development, implementation, and operation to increase patient engagement during hospitalization. Quantitative survey data can be analyzed using SPSS for descriptive statistics and qualitative data using NVivo software for themes. It is encouraged that such surveys are conducted with other institutions to support this area of research. Results from analyzed surveys will be used to build a conceptual model for acute care patient portals and make recommendations for future investigation.

Acknowledgments: This BWH Libretto Task Force is funded by the Libretto Consortium supported by the Gordon and Betty Moore Foundation. We acknowledge the Libretto Task Force team members for their contribution to and feedback for the survey questions.

References
A Deep Learning Framework for Improving Medical Information Retrieval
Ding-Cheng Li PhD1, Saeed Mehrabi, PhD1, Yue Yu, MS1,2, Honghua Liu, PhD1
1Mayo Clinic, Rochester, MN; 2Jilin University, Changchun, China

Background and Introduction. Many innovations in language modeling for information retrieval can be considered to combine multiple alternative document representations or “layers” [1]. However, questions incurred are what layers should be included and whether we can find layers which can be general enough to fit all kind of data. One common agreement is that concepts are playing gaining roles in building more advanced language models. Zhu and Carterette represent a document in separate text and concept spaces and then combine them into a statistical language model. Wei and Croft represent a document via latent Dirichlet allocations. Both of them reported higher performances were achieved. Recently, deep learning evolved from traditional neural networks have seen widespread use in natural language processing due to its powerful capacities of catching hidden similarities among words and phrases. However, up to now, no system has been constructed with deep learning for medical information retrieval. In this work, we propose to build a new layer with deep learning in order to empower our existing layered language model, where bag of words, medical concept unique identifiers and dependency relations are integrated.

Methods. Based on previous work, our proposed framework focuses on two steps, associating each word w with a vector representation \( x_w \in \mathbb{R}^d \) and computing distributed representations for each sentence with a dependency-tree recursive neural network (DT-RNN) [2]. Word-embedding is a technique to represent individual words with a vector. As we know, words themselves are often incomplete in meaning and full of ambiguous. They have different meanings under different contexts. Word-embedding overcomes this weakness by making use of the morphological and syntactic knowledge of words in a sentence or a passage or some slides of windows in a context to transform each of them into a rich semantic representation. We employ GloVe (global vectors for word representation) model to implement the word-embedding. These vectors are stored as the columns of a \( d \times V \) dimensional word embedding matrix \( W_e \), where \( V \) is the size of the vocabulary and \( d \) is the length of the dependency tree. DT-RNN then takes dependency parse trees (obtained with Mayo eTAKES pipeline) of both queries and the medical records as input.

Each node \( n \) in the parse tree for a particular sentence is associated with a word \( w \), and accordingly a word vector \( x_w \). Unlike in constituent trees, where all words reside at the leaf level, the internal node in the parse tree is a hidden vector, which are associated with words and arcs connecting nodes represent the dependency relations among words. Taking advantages of information of the hidden vector, DT-RNN aims at optimizing the semantic representation of the corresponding words, phrases or sentences. Specifically, it combines the each node’s word vector with its children’s hidden vectors to form \( h_n \). This process continues recursively up to the root of each dependency tree, namely, the whole sentence. The combination takes the sigmoid function as \( f(W_{e} \cdot x_{w} + b + \sum_{k \in K(n)} W_{R(n,k)} \cdot h_{k}) \) where \( W_{e} \) is a \( d \times d \) matrix with each dependency relation \( r \) in our dataset (similarly, the matrix is built with GloVe) and \( R(n,k) \) is the dependency relation between node \( n \) and child node \( k \).

Experimental Setup and Initial Results. Both 2011 TREC-med queries and 2012 TREC-med queries are utilized for the experiments. They are free text of medical records provided by the University of Pittsburgh’s NLP Group. Each patient at the University of Pittsburg has one more medical records associated with him or herself. Each record contains both structured and unstructured data in XML format, with an average length of 423 words. The overall vocabulary size is on the order of \( 10^5 \). The notes were de-identified, so that any protected health information has been replaced with surrogates. In total, there are 95,702 records that correspond to 17,198 visits. Both records include 81 hypothetical queries that might be used to identify participants. Up to now, we have successfully generated vectors for all the vocabularies and algorithms of the second step has been implemented. After we obtain results of this layer, we will integrate the scores into our layered language models with assigned weights (like other layers, the weight parameter \( \lambda \) and the smoothing parameters \( \mu \) are learned with grid search). Then, the MAP performance of retrieval will be reported.

Acknowledgements
The study was supported in part by the following grants: R01GM102282A1, R01LM11934A1, R01LM11829A1, and R01LM11369A1.


1554
A survey of automated information retrieval for genetic disorder from GeneReviews

Pu Li, MS¹, Qian Zhu, PhD¹

Department of Information Systems, University of Maryland Baltimore County, Baltimore, Maryland, USA

Introduction.
A genetic disorder is an illness caused by one or more abnormalities in the genome, especially a condition that is present from birth (congenital). Most genetic disorders are quite rare and affect one person in every several thousands or millions. That is to say not too much information about the genetic disorder is available for supporting relevant research. In another hand, GeneReviews is expert-authored, peer-reviewed disease descriptions ("chapters") presented in a standardized format and focused on clinically relevant and medically actionable information on the diagnosis, management, and genetic counseling of patients and families with specific inherited conditions. GeneReviews in free text can be browsed and searched via NCBI website, however, no structured electronic version available to be integrated into the clinical decision support systems, especially for the Genetic Testing Ontology (GTO) developed in our previous work¹². In this study, we conducted a survey by applying two widely-used biomedical annotation tools, National Center for Biomedical Ontology (NCBO) annotator and Unified Medical Language System (UMLS) MetaMap to annotate data extracted from 10 GeneReviews chapters. From this preliminary work, we identified gaps between a large volume of genetic information and insufficient capability of processing such information by using current Natural Language Processing (NLP) tools, and propose the extension of these tools to be able to annotate relevant genetic information.

Methods.
10 GeneReviews chapters corresponding to 10 different genetic disorder have been randomly selected and downloaded from NCBI (http://www.ncbi.nlm.nih.gov/books/NBK1116/). Each chapter includes 10 sections, Summary, Diagnosis, Clinical Description, Differential Diagnosis, Management, Genetic Counseling, Resources, Molecular Genetics, References, and Chapter Notes. In this study, we primarily focused on 7 sections, except for Resources, References and Chapter Notes. Each of the selected sections has been programmatically extracted and annotated by NCBO annotator and MetaMap via their APIs. We manually reviewed and evaluated the results to identify the coverage and gaps of these two NLP tools.

Results.
Total 50,687 annotations have been generated by NCBO annotator and 8,479 annotations by MetaMap. We manually reviewed all annotation results comparing to the original chapters, the average accuracy rate is 74.04% for NCBO annotator and 85.23% for MetaMap. Although 8,749 annotations are exactly matched with UMLS concepts with mapping score equaling 1000, there are some unnecessary mappings, such as “treatment”, “syndrome” that were grouped into incorrect mappings.

Discussion.
This is a preliminary work of examining the feasibility of extracting genetic disorder related information by using the current NLP tools being widely used in medical field. The results produced from this survey not only illustrates that the majority of the information, especially for medication, diagnosis relevant information can be annotated with high accuracy, but also guides us to propose extensions of the existing NLP tools to be able to manage genetic information accordingly based on the gaps identified. The majority of missing/incorrect annotations we identified are falling two categories, genomics and numerical values with relevant context, which are both critical information for genetic disease management. For instance, “Each sib of a proband with 3-M syndrome has a 25% chance of being affected, a 50% chance of being an asymptomatic carrier, and a 25% chance of being unaffected and not a carrier.” Such information about the percentage along with relevant context is critical to predict the risk of affecting 3-M syndrome. Thus, either manual annotation or appropriate component development will be the next step to integrate a complete list of annotations from GeneReviews into the GTO to support iGenetics³.

Reference.
Detecting Mitral Valve Prolapse (MVP) From Heart Sound Recordings

Yihua Li, Dorothy W. Curtis, M.Sc.
Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA

Overview
In this project, we apply machine learning techniques to detect Mitral Valve Prolapse (MVP) from phonocardiogram (PCG) and electrocardiography (EKG) recordings of the heart. After segmenting each recording into individual heart beats, we use Gaussian mixture model to assign the beats into disease and normal clusters based on the presence of murmurs. Features are extracted from this beat level clustering and are used to classify each patient with support vector machine (SVM). We achieve an overall precision rate of 94.7% and recall rate of 88.9% on our dataset. This automatic heart sound analysis is valuable in the diagnosis process and should reduce false positives.

Introduction
MVP is a cardiac disorder characterized by the displacement of an abnormally thickened mitral valve leaflet into the left atrium during systole. During cardiac auscultation, skilled cardiologists detect MVP by the presence of murmurs, high frequency noise-like sounds caused by abnormal blood flow patterns. However, the interpretation of heart sounds is a difficult skill and subject to human auditory limitations. As a result, automatic MVP detection would be very helpful. Previous works focus on the classification of different heart valve diseases by machine learning methods [1]. Here, we focus on the case of MVP and seek to integrate beat level classification into the final decision, considering that not all beats in the recordings of a MVP patient exhibit murmurs.

Dataset Description
The dataset, provided by Massachusetts General Hospital, comes from 43 healthy individuals and 36 MVP patients. For each subject, EKG and PCG were recorded from 1 to 3 different locations of the heart. Each recording is roughly 30 seconds long. We give the recordings the same label as that of each person given by the physician. Half of the data is used as the training set and half as the testing set.

Method
We first carry out classification on the beat level. The PCG recordings are segmented into heart beats through peaks in the EKG recordings. The locations of s1 and s2 are identified as in [2]. We then take the wavelet transform of each beat using a Daubechies wavelet with 2 vanishing moments, and perform a level-8 decomposition of the signal. This wavelet transform allows us to separate murmurs from baseline sounds. For the features, we look at the decomposition of the beat at level 7. Since the disease beats typically have high energy murmurs during the systolic phase, the chosen features are: (1) the energy during systole normalized by the energy during s2 and (2) the location of the peak during systole. With these features, we then soft assign the individual beats from all recordings into disease and normal clusters using Gaussian mixture models. We apply this unsupervised method since only the labels for recordings, not individual beats, are available and a MVP recording could contain normal beats. All the beats from normal recordings and a small subset of manually labeled disease beats are used to initialize the clusters. Next, we classify whether a patient has MVP using SVM. To obtain the features, for each recording we calculate the following: (3) the percentage of disease beats, (4) the posterior probability that a beat belongs to the disease cluster, averaged across all disease beats in that recording. As a patient has multiple recordings, the final features for each patient are the average and the maximum of (3) and (4) across his/her heart sound recordings.

Results
We analyze the data in MATLAB and achieve an overall precision rate of 94.7% and recall rate of 88.9% on our testing set. Our future work is to take into account, during classification, the location where the recording was taken.

Conclusion
We designed a 2-level classification method for automatically detecting MVP from heart sound recordings. The preliminary result on our dataset suggests that our method has the potential of being applied to real-life diagnosis.

References

1556
Design of a “Synthetic” Data Set for Teaching and Evaluating Analytics Methodology in Accountable Care Organizations

Robert W. Ligon, MS,
University of Texas, Austin, Texas

Abstract

See submission content

Introduction

Accountable Care Organizations (ACOs) were introduced in the Affordable Care Act as a pilot program developed by CMS to pilot shared savings as a model of assuring quality while reducing costs in healthcare delivery. The makeup of these organizations can vary; but they often involve multiple healthcare organizations including physicians, hospitals, clinics, ambulatory surgery centers, laboratories and pharmacies. These distributed data environments create several challenges for ACO data governance.

First is data standardization. Often the participating organizations have different EHR products with different data requirements, data standards and data definitions. Secondly, is the creation of a master patient index to identify patients throughout the continuum of care. Thirdly a centralized system for providing data standardization and preprocessing and finally a system that will allow for analysis of clinical, financial and outcomes data for reporting.

When presented with these challenges many organizations lack sufficient training to understand the basic requirements for ACO data management. The purpose of this study was to create a mock ACO data environment using currently existing data as well as modeled “synthetic” data environments to simulate the data needs and data challenges of a multi-organizational ACO.

Methods

We used currently existing de-identified hospital discharge data from the state of Texas to create two fictitious hospitals each with one year discharge rates of approximately 20,000 patients. Fictitious patient identifying information was created to provide patient names, date of birth, dates of service and unique hospital medical record number. Data for each hospital included diagnoses, procedures, outcomes and detailed charge and reimbursement information. Outcome information was derived from provided ICD-9 codes for in-hospital events. Risk algorithms were created using available demographic and diagnostic information.

Physician practice data was modeled using simplified EHR data structure and included patient demographics, encounter level data including diagnosis, procedures and charges. Charge and reimbursement information was simulated using Medicare fee schedules as a percent of private pay reimbursement. Meaningful use clinical and process measures were simulated. Clinical data was modeled using published incidence and prevalence information when available. Physician panels were set to between 1800 and 2200 patients. The created physician practices included two primary care, one OB/Gyn, one cardiology and one orthopedic practice. There was no master patient identifier between organizations.

Conclusions

This “synthetic” ACO environment allows users to query relatively straight forward information on utilization, clinical measures and outcomes but also allows for users to create a simple master patient index, evaluate inter-organization utilization and evaluate outcomes, capitated contracts over time and longitudinal cost of care.
Designing iSee, the intelligent Search expansion tool
Ko-Wei Lin, DVM, PhD, Dexter Friedman, Hyeoneui Kim, RN, MPH, PhD
Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA

Introduction and Background
Vast amounts of biomedical data became available, which we can reuse and assimilate to evaluate existing hypotheses and to make new scientific discoveries. Data abundance also poses new challenges, particularly in terms of the efficiency and accuracy of searching through this data, due to idiosyncrasies in its representation. It is often mentioned that a recommendation system would be potentially useful in facilitating the search of biomedical data. For example, the biomedical researchers we interviewed to collect user requirements for PhenDisco (http://phendisco.ucsd.edu), a concept-based search tool we built for dbGaP (http://www.ncbi.nlm.nih.gov/gap), suggested that we implement automatic search expansion (i.e. a “recommendation” feature) within PhenDisco. Unlike in commercial domains, such functionality has mostly been developed and used in an experimental setting in biomedical research.

bioCADDIE (biomedical and healthCAre Data Discovery and Indexing Ecosystem, http://biocaddie.org) is a project that aims to facilitate the discovery and indexing of biomedical data in order to support large scale data analysis and reuse. iSee (intelligent Search expansion) is a pilot project funded by the bioCADDIE initiative. The main goal of iSee is to support efficient data search and review by intuitively expanding the user’s search to return highly accurate records in an easily assessable manner. iSee is to be developed as a new feature in PhenDisco. We are piloting this idea with PhenDisco as it already has search functions that iSee can expand and it contains annotated phenotype data with standardized metadata. iSee development involves two parts: (1) generating record relevancy scores based on readily available standardized metadata and (2) designing an intuitive user interface. This work reports on the lessons learned during the initial requirement-gathering phase for the latter.

Methods
We reviewed 10 web-based biomedical data repositories and 5 commercial sites with data search capabilities to see how the automatic search expansion function is implemented. We also interviewed three biomedical researchers who are familiar with genome-wide association studies or genomics databases. During the interviews, we first gave a 10-minute presentation on the background and goal of iSee development, and provided a 3-minute demonstration with a straw-man iSee model to convey the iSee idea more clearly. iSee straw-man contained a study map in which relevant studies were arranged by metadata parameter similarities, a customized parameter selection section allowing users to prioritize the metadata relevant to their search, and a study information section. The goal of these interviews was to learn (1) realistic workflows around search and search expansion, (2) types of information that search expansion algorithms need to take as high priority, and (3) considerations on user interface designs.

Lessons Learned and Future Work
All biomedical websites we reviewed support an advanced search menu where users can refine search criteria. However, only two of them\(^1\) provide a function similar to automatic search expansion or recommendation, wherein the tool shows additional records comparable to the returned record of interest. These sites utilized metadata associated with the records/data to generate various views, thereby expanding its search further to other related records. The commercial websites all provide a recommendation system largely based on the users’ transactional log data, which was deemed less relevant to current iSee development.

The input from the biomedical researchers can be summarized as follows: (1) produce a view with comparable studies (i.e., expanded search returns) based on the default parameter setting, (2) allow users to refine/manipulate the comparable study view using the full set of available metadata, (3) allow users to set their own relevance criteria by assigning different weight to each metadata item, (4) an interactive graphical display is preferred but follow conventions to minimize the cognitive burden to understand the symbols and notations, and (5) although a bit cluttered, having more complete information available in one view is preferred to requiring additional user actions (e.g., mouse-over or clicking) to get to the necessary information. A revised iSee straw-man model is to be released in August of 2015. We plan to implant parameter similarity algorithm and metadata hierarchy in iSee system, and reevaluate the new model with a larger group of potential users.

Acknowledgment This work was supported by the NIH BD2K grant 1U24AI117966-01. We thank Drs. Melissa Haendel and Caroline Nievergelt, and Mr. Adam Maihofer for providing invaluable insights on iSee development.

\(^1\) The full summary table is available at https://idash-data.ucsd.edu/download/item/131955/Repository+Comparison.xlsx
Health Care Providers’ Perceived & Actual Problems in the Use of HIT in the ED

Mary L. Little, RN, MSN1, Osman Sayan, MD2, Vimla L. Patel, PhD1,3
1Dept of Biomedical Informatics, Columbia University; 2Dept of Emergency Medicine, Columbia University; 3New York Academy of Medicine, New York, NY

Abstract
Emergency Departments (EDs) are complex environments due to the interdependent, episodic, collaborative, nonlinear, and decision dense nature of the patient care.1 To gain a true perspective of provider perceptions and the reality of the ED workflow environment, recorded semi-structured interviews, ethnographic observations, recordings of patient handoff information, and digital sound pressure level (SPL) data were collected. The preliminary analysis indicates a mismatch between healthcare providers’ focus and perceptions of the seriousness of electronic health record (EHR) navigational difficulties, while ignoring the actual impact of task interruption and excessive noise levels on cognitive functioning. A discrepancy between providers’ perceptions and reality, suggests there may be a risk when health information technology (HIT) system changes are performed on interview data alone. Our findings suggest ethnographic observations in addition to interviews may be necessary to obtain a complete picture of an ED workflow environment prior to implementing any HIT system alterations.

Introduction
The ED environment is characterized by EHR time constraints and the need for urgency, information immediacy, rapid task completion, and task accuracy. If technology does not facilitate these activities, we suspect a mismatch between perception and reality will occur as providers focus on HIT interface navigation issues and overlook workflow impairment factors such as task interruptions and excessive noise levels which negatively impact cognitive functioning.2 Assessing the accuracy of provider perceptions and the reality of the healthcare environment is critical to effective HIT system alterations.

Methods
Nine detailed semi-structured interviews of ED resident and attending physicians were recorded to capture perceptions of HIT use in the ED, care coordination, and care quality and delivery. Fifteen care transition clinical rounds were recorded. All recordings were digitally encrypted, with patient health information (PHI) removed prior to transcription. Additionally, over 30 hours of digital SPL recordings were conducted in conjunction with ethnographic physician patient handoff and workflow task observations. Interview responses were coded using NVivo software,3 observations and patient handoff audio-recordings were analyzed separately. SPL data were analyzed by professional sound engineers and combined in a preliminary results analysis.

Results
Interview results revealed provider perception of the difficulties of working with HIT as being the most serious problem in workflow and task completion. While ethnographic observations revealed frequent interruptions by staff, consulting physicians, patients, and families, which often occurred during critical patient handoff information transfer. Noise levels, especially during high capacity time periods were much higher than anticipated with peak noise levels often greater than 90 decibels.

Discussion
Healthcare providers’ perceptions of the seriousness of EHR navigational difficulties overshadows the reality of the seriousness of task interruptions and high noise levels in the workflow environment. It is likely that providers’ HIT interface needs must be satisfied before attention to causes of cognitive impairment can be recognized or addressed.

Conclusion
Providers may perceive difficulties of working with technology a more serious problem than workflow issues affecting cognitive functioning. A mismatch in provider perception and the reality of the healthcare environment underscores the importance for researchers to gather ethnographic data or other forms of real life observations to assess the validity of healthcare providers’ perceived problems prior to HIT system alterations.

References
1. Plsek PE, Greenhalgh T. The challenge of complexity in health care 2001-09-15 07:00:00. 625-8 p.

Acknowledgment: Funded by 5 T15 LM007079. Research funded by James S. McDonnell Foundation (Grant No. 220020152 to Vimla L. Patel). The authors thank James Giglio, MD for his assistance and Sana Khalid, MS, for her contributions to the ethnographic observations.
An analysis of PubMed4Hh App User Distribution
Fang Liu, MS, Paul Fontelo, MD, MPH
National Library of Medicine, Bethesda, MD

Abstract
PubMed for Handhelds (PubMed4Hh) is an app for finding relevant health information from the National Library of Medicine on mobile devices. Apple’s iOS app developer tool provides daily downloads data and regional distribution data. Comparison between PubMed4Hh download distribution and the regions of PubMed citations shows a consistent match between the number of users and the number of PubMed indexed publications of a region.

Introduction
The PubMed4Hh app has been on Apple iTunes store since July 2013. The app searches PubMed citations with multiple search strategies1. We observed consistently that 55% of PubMed4Hh app downloads are from outside the continental United States. At the end of 2014, more than 230K downloads were recorded for iOS devices.

Methods
Apple’s iTunes Connect provides daily and yearly download data to applications’ developers including data on regions and locations. PubMed statistical reports show the countries and the number of citations from these countries’ grants respectively2. From the PubMed database, the number of citations for multiple languages can be retrieved by setting the different language filters. We analyzed the data from iOS downloads from the iTunes and statistical reports of the most PubMed grant countries from the PubMed citations database.

Result

<table>
<thead>
<tr>
<th>Number of Downloads</th>
<th>Number of Countries</th>
<th>Top 10 Country</th>
<th>Total PubMed4Hh Downloads</th>
<th>% of Total</th>
<th>Language</th>
<th>Number of PubMed citations</th>
<th>PubMed Grant Countries</th>
<th>Number of Citations with Grants</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;10</td>
<td>26</td>
<td>US</td>
<td>105765</td>
<td>45.90%</td>
<td>English</td>
<td>20191764</td>
<td>USA</td>
<td>4,329,260</td>
</tr>
<tr>
<td>10-99</td>
<td>43</td>
<td>BR</td>
<td>13945</td>
<td>5.93%</td>
<td>Portuguese</td>
<td>89388</td>
<td>UK</td>
<td>158,222</td>
</tr>
<tr>
<td>100-999</td>
<td>24</td>
<td>IT</td>
<td>9227</td>
<td>3.93%</td>
<td>Italian</td>
<td>295425</td>
<td>Italy</td>
<td>8,811</td>
</tr>
<tr>
<td>1000-10K</td>
<td>29</td>
<td>DE</td>
<td>7965</td>
<td>3.39%</td>
<td>German</td>
<td>815331</td>
<td>Canada</td>
<td>34,085</td>
</tr>
<tr>
<td>&gt;100K</td>
<td>1</td>
<td>ES</td>
<td>7139</td>
<td>3.04%</td>
<td>Spanish</td>
<td>308177</td>
<td>International</td>
<td>2,656</td>
</tr>
<tr>
<td>Total</td>
<td>124</td>
<td>CA</td>
<td>6791</td>
<td>2.89%</td>
<td>Chinese</td>
<td>255613</td>
<td>Ireland</td>
<td>700</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CN</td>
<td>6510</td>
<td>2.77%</td>
<td>Dutch</td>
<td>63091</td>
<td>USA</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 1. PubMed4Hh app download by country; PubMed Citation in multiple languages and PubMed grant countries

PubMed4Hh app was downloaded by users from 124 countries. In 31 countries (25.0%), there were more than 1000 copies of the app downloaded and in 24 countries (19.3%), more than 100 but less than 1000 apps were downloaded (Table 1). Languages spoken in the top 10 downloading countries are English, Spanish, Portuguese, Italian, German, Chinese and Dutch. Four PubMed Grant countries (US, UK, IT and CA) had the highest downloads. From the PubMed citations database, languages used by the most downloading countries all have large numbers of citations of that language ranging from 63K to 20M.

Conclusion
Data analysis shows the countries that had the highest PubMed4Hh app downloads all have a significant number of citations in PubMed. The topmost PubMed grant countries were also countries that had the most PubMed4Hh app downloads.

References
Caveats of Using Social Media Data for Medical Research: A Report from a Study on Eye-Related Symptoms in Tweets

Yang Liu MSJ,1 Tricia O’Brien,2 Esha Sondhi,3 Qiaozhu Mei PhD,1,4 David A. Hanauer MD, MS,5,1 Kai Zheng PhD6,1
1School of Information; 2College of Literature, Science, and the Arts, 3College of Engineering; 4Department of Electronic Engineering and Computer Science; 5Department of Pediatrics; 6School of Public Health Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA

Introduction
Social media websites are increasingly used by the general public as a venue to express health concerns and disseminate information during public health crises. Numerous prior studies have demonstrated that social media can be a valuable source of information for purposes such as early detection of disease epidemics and clinical and public health research.1 Nevertheless, researchers have also cautioned about the use of social media data because such data, which could be generated by anybody on the internet, are inherently noisy.2 In this study, we applied MetaMap to a random sample of twitter messages to identify potential mentions of eye-related sign/symptom concepts. We then conducted a manual review of the results extracted by MetaMap to assess their medical relevance. The objective was to identify non-medically relevant tweets that may be ‘mislabeled’ by MetaMap, and common reasons for this mislabeling.

Methods
We analyzed a 10% random sample of all geo-tagged English tweets generated in 2013 collected through the Twitter stream API with Gardenhose access. Clinical concepts in the tweets were extracted using the natural language processing tool MetaMap.3 Two coders then manually reviewed a random sample of the MetaMap-produced results to identify tweets that were not medically relevant. In this process, the two coders also classified the non-medically relevant tweets into categories that best characterize the reasons for their being mislabeled.

Results
Out of approximately 300 million tweets analyzed, MetaMap identified 291 distinct types of “signs” and “symptoms.” 16 of them were considered as eye related by two ophthalmologists (e.g., bloodshot eye, blurred vision, redness of eye). These 16 eye-related signs and symptoms appeared in 25,883 tweets, 1,555 of which were manually reviewed till saturation was achieved (i.e., no new categories continued to emerge from the data). Out of these 1,555 tweets, the two coders agreed on 1,387 of them regarding whether they were medical relevant. The inter-rater agreement rate (Cohen’s kappa) is 0.76.
Among these 1,387 tweets, 960 (69.2%) were deemed as non-medically relevant by both coders. The most common reasons for mislabeling include “different meaning” (39.4%), “lyrics/poem/quote” (19.5%), “emoticon” (12.9%), “words out of order” (8.3%), “circumstantial” (7.6%), and “idiomatic” (7.3%).

Conclusion
In this study, we demonstrate that MetaMap, a popularly used medical named-entity extraction tool, performed very poorly in extracting medical concepts from tweets, a common type of social media data. Of the tweets identified by MetaMap as containing medical relevant concepts, 70% were deemed by the two coders as non-medically relevant. Extreme caution shall be therefore used when applying NLP tools to extract medically relevant content from social media for disease surveillance or for research.

References
A Cross Strait Cooperation Design and Implementation of Mobile Nursing Information System

Yun Liu, MD1, Pinjen Huang, MS, RN2, Zejuan Gu, RN3, Rong Wang, RN3, Chunhong Gao, RN3, Tao Shan1, Jianjun Guo1, Zhongmin Wang1

1Jiangsu Province Hospital Information Department, Nanjing, China; 2MiWell, Taipei, Taiwan; 3Jiangsu Province Hospital Nursing Department, Nanjing, China

Introduction

Under China’s 12th Five-Year plan, hospitals need to adopt EHR, change nursing care model from functional care to primary care, and manage nursing workload in a scientific way. Based on above guideline, we evaluated current nursing information systems in China, and could not find any of them meets the requirement. Our original nursing information system is an add-on function to health information system (HIS) and each nursing ward would be equipped with 1-2 desktops for nurses to access HIS. Nurses have to print out orders on papers daily, or whenever physicians update their orders. After giving medication, nurses need to sign the paper forms; those forms would be kept for 14 days and then discarded after patient discharged. Taiwan has rich ICT resources and the development of nursing information systems there is sophisticated and relatively mature. We consulted Taiwan’s nursing informatics experts and identified a start up company with a prototype of mobile medication administration record system (mMAR)2 which was implemented with a qualified design concept. The company was invited to cooperate with us to redesign a new mobile nursing information system (mNIS). The implementation of the system was separated into two phases: phase one is order execution associated nursing process, and phase two is nursing plan and nursing documentation. In this study, we will share our phase one experience from design to implementation.

Materials and Method

This study was conducted in a more than 3,000-bed medical center in Nanjing city. The mNIS committee was setup in Oct 2013, led by nursing deputy director. The members include head nurses, a nursing informatics PM, an engineering PM, an user experience designer and engineers. During process redesign, physician affair director, laboratory director, blood bank director and pharmacy director joined every process discussion. We implemented barcode-printing function in pharmacy intravenous admixture services (PIVAS) so nurses don’t have to do the medication label cutting and pasting. During the implementation of mNIS, hospital HIS rules were also clarified. Physician’s prescription behaviors were therefore standardized and nurses don’t have to judge the first dose of medication purely based on experience. All information associated with medication administration, blood transfusion and blood sampling was integrated into one system, which used to be separated in different portals. We chose MioCARE A105, an Android 4.0 tablet 6” display and integrated with 2D barcode scanner as our main system interface. C++, Java were used for developing the native app, task server, and sync server, while PHP was used for web portal implementation. SOAP and Socket were used for exchanging data with each hospital’s information system. Phase 1 contains 3 process (medication administration, blood transfusion and blood sampling) and 9 modules. Two geriatric wards were selected as power users, they were trained and feedback to mNIS committee module by module from August to November 2014. From mid of Dec 2014 to Mar 2015, power users were asked to use paper forms and mNIS in parallel. A 5-point Likert scale questionnaire designed by Technology Acceptance Model for Mobile Service (TAMM) model was used to evaluate nurses’ acceptance of mNIS. There were 14 nurses invited to answer the questionnaire at the first week of Mar 2015.

Result and Conclusion

The highest score of questionnaire was Ease to learn (4.24 out of 5) follow by Ease to use (4.16), Usefulness (3.73), Trust (3.72), Acceptance (3.7), and Satisfaction (3.55). The questionnaire indicated a positive feedback from nurses even with some extra loading caused by the request having mNIS and paperwork processes running in parallel. We will optimize the phase one performance then go live on 60 wards separately before the end of 2015.

References

1. Feng RC, Yeh YT. A new vision of nursing: The evolution and development of nursing informatics. The Journal of Nursing. 2014;61(4, Suppl.):78-84
Developing Clinical Decision Support for Patient Self-Management: A Prototype for Symptom Management in Cancer Patients

David F. Lobach, MD, PhD, MS1; Janet L. Abraham, MD2; Donna L. Berry, RN, PhD2; Michael S. Rabin, MD2; Ilana Braun, MD2; Manan Nayak, MA2; Mary E. Cooley, RN, PhD2
1Klesis Healthcare and Duke University Medical Center, Durham, NC; and 2Dana-Farber Cancer Institute, Boston, MA

Abstract
Patient-centered care includes enabling patient self-management. Improved clinical decision support (CDS) tools designed for direct use by patients are needed to empower patients to self-manage their symptoms. We used a formative process to design a prototype CDS tool for self-management of symptoms by cancer patients. This prototype reflects design principles created to address patient barriers and clinician concerns for patient use of CDS. Patient usability surveys indicated strong support for the CDS tool.

Introduction
Relatively few rule-based CDS tools have been developed to directly assist cancer patients with symptom self-management.1-4 In those studies, the focus of the CDS was rudimentary1, relied heavily on interactions with clinical personnel2,3, or was implemented as a subcomponent of more comprehensive cancer information system.4 The goal of this formative study was to identify and create a prototype CDS tool that overcame identified patient barriers and clinician concerns regarding the direct use of CDS by cancer patients for symptom self-management.

Methods
This project focused on the most common symptoms contributing to emergency and urgent care visits among patients with cancer, namely pain, constipation, and nausea and vomiting. Our multi-disciplinary research team conducted a formative mixed methods study involving focus groups, interview sessions, and usability surveys with cancer patients, their caregivers and oncology clinicians. All focus group and interview discussions were audio-recorded and content was transcribed for analysis. Based on the data gathered and on input from stakeholders participating on expert panels that included patients, caregivers, clinicians, information system developers, and administrators, we identified patient barriers and clinician concerns regarding the use of CDS by cancer patients for symptom self-management. Addressing these barriers and concerns led to design principles that informed CDS tool prototypes to simulate design and content features of a self-management CDS system for patient symptoms to which study subjects could respond.

Results
We recruited 24 patients and 13 oncology clinicians to participate in the formative evaluation. CDS design principles addressed patient safety, cultural competency, care coordination, resource availability, and system function. Specific features of the prototype included identification of emergency issues, specific details for when and how to contact the clinical team, access to explanatory information, and expedited data collection to place users at the appropriate place within the CDS algorithm. Patient usability surveys indicated strong support for the CDS tools.

Conclusions
The findings from this study provide insight into how CDS systems can be developed to support patient self-management directly. Patient safety and tool navigation were critical features. Many of the insights gleaned from this study may be generalizable and could inform the development of CDS resources that patients with other diseases and their caregivers can use for self-management of symptoms.

Acknowledgements
This project was funded in part by The Patient Centered Outcomes Research Institute Grant PI-12-001.

References
Real-time SNOMED Post-coordination of Adverse Drug Reactions: Model Formulation for an Actionable Registry

Marcelo Lopetegui, MD, MS¹,², Mario Barbé, MD¹, Daniel Luna³, Alejandro Mauro, MD¹
¹ Clínica Alemana de Santiago, Facultad de Medicina, Universidad del Desarrollo, Santiago, Chile
² Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA
³ Health Informatics Department, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina

Introduction
Prevalence of allergies has increased in the last 50 years: currently 40-50% of the population presents some level of sensitization, while 10% will present an adverse drug reaction¹. Two percent may present an episode of anaphylaxis in their lifetime, with a reported lethality of up to 2%.² Thus, updated and accurate allergies records are key for a safe patient care. Furthermore, appropriate allergy records in interoperable environments could decrease medication related errors³. In this context, we aimed at developing an allergies registry module in our home-grown Electronic Health Record (EHR) attempting to achieve two main outcomes: maintain updated allergic status for our patients, and do it on a manner that produces actionable data we could leverage, all while being minimally disruptive in the usual clinical workflow.

Methods
Updated and meaningful allergic status. Based on user-centered design, we defined four allergic statuses: unknown, positive, negative, and outdated. The allergic status is always visible in the patient header, allowing for a quick and meaningful assessment of the allergic status of the patient. The unknown status (never asked) is expected to provide an evident differentiation from a negative status (answered no), motivating the clinician to assess the patient’s initial allergic status. In order to ensure updated data, we configured the module to consider the status valid for one year. Following international recommendations, we decided to require the allergy status to be up-to-date in order to prescribe a new medication in the Computerized Physician Order Entry (CPOE).

Fast data entry and actionable data. Text fields provide fast data entry, but are hard to computationally act and reason upon, while simple lists provide limited sets of structured data. We desired fast data entry without compromising the quality and usefulness of the registered data. Thus, we developed a real-time concepts post-coordination module, which, based on SNOMED CT, permits the clinician to easily enter any potential allergic status, identifying the causing agent (drug, food or environmental agent), the symptoms and signs, and the severity. The result is a completely customized list of allergies for each patient, recorded with all the appropriate relationship types to existing SNOMED CT concepts.

Results
A first version of the module was deployed in our emergency department information system for verification purposes, currently ongoing stability upgrades. The evaluation of the proposed model will be conducted in the outpatient setting based on a pre/post assessment of the proportion of patients with a reported allergic status, the proportion of patients with statuses updated within the last year, the assessment of the perception of the module by clinicians, and with a feasibility study of cohort creation based on detailed allergic statuses.

Conclusion
We present a model to manage the registry of adverse drug reactions and allergies in a home-grown EHR, providing an actionable registry based on post-coordination of standardized concepts, and expecting to improve allergic status completion and validity. Ongoing assessment will be presented in a future report.

References
Mapping Workflows in a Surgical Clinic to Guide Implementation of a Patient-Centered Postoperative mHealth Wound Assessment System

Ross Lordon, BS1, Heather Evans, MD1, Andrea Hartzler, PhD2, Cheryl Armstrong, BSN RN MPH1, Sarah Whitehead, ARNP MN1, Patrick Sanger, PhD1, William Lober, MD1

1University of Washington, Seattle, Washington; 2Group Health Research Institute, Seattle, Washington

Intro: The successful adoption of novel health information technology (HIT) is often challenging and susceptible to a multitude of problems. Poor design, lack of planning, ineffective implementation, or an absence of socio-technical support post-implementation, can cause technology to be underutilized or abandoned entirely. Effective implementation demands a comprehensive understanding of the current work processes of the target environment. Once established, it is imperative to closely align new technology with current work processes to minimize frustration during implementation1. Anticipating user needs, such as nurses integrating patient generated mobile health (mHealth) data with the electronic health record (EHR), for implementation is paramount. Specifically, our team is implementing an mHealth platform to facilitate patient self-tracking and communication with providers about the signs and symptoms of surgical site infections after hospital discharge. Patients utilize an mHealth application and providers access the patient generated data on a computer based dashboard. Unfortunately, we lack guidance on workflow mapping for implementation of mHealth platforms in clinical settings. The majority of guidance on HIT workflows focuses on EHRs and decision support, such as alerts and reminders. To explore this gap, we examined the adaptation of existing HIT workflow assessment tools recommended for EHRs, such as benchmarking and flowcharting,3 to prepare for the implementation of an mHealth platform in surgery care. Methods: We conducted semi-structured interviews with staff in a plastic surgery clinic to map the existing workflows for utilizing wound photos submitted by post-operative patients. We formally documented these workflow processes using business process-modeling notation (BPMN), a form of flowcharting.3 Once the existing workflows had been visualized in the first BPMN diagram, we produced a second BPMN diagram depicting the proposed changes to the existing workflow. We presented the second BPMN diagram to all the stakeholders involved with the mHealth platform implementation. Barriers, opportunities, and weaknesses were identified and included in subsequent revisions. Results: After three rounds of iterative revisions, the BPMN diagram helped us to align the implementation of the mHealth platform with current clinic work processes. Next we used our findings to draft a new BPMN diagram depicting workflow processes after future implementation of the mHealth platform. Then we presented this to the clinic and mHealth development team to guide the implementation process. Using this diagram, the mHealth development team and clinic staff identified specific components of the workflows that will provide evaluation metrics to assess process improvements after implementation. Examples of metrics include, nurses’ time spent discussing wound concerns with patients, time spent documenting concerns in the EHR, and time to resolve reported problems. Discussion: Generating BPMN diagrams demonstrating clinic workflow processes before and after the mHealth platform implementation helped us gain the confidence of the clinic staff, streamline the implementation planning process, and lay down a critical foundation for the implementation of our mHealth platform in clinical care. Through this iterative process, we gained insights into the multiple ways a patient could initiate communication about wound concerns. We incorporated these insights into new workflows and evaluation metrics to facilitate future implementations. Thus, the BPMN diagrams helped inform the new workflow processes with the intent of minimizing disruption to the current workflow within the clinic. The subsequent implementation phase will incorporate producing a third BPMN diagram to depict the actual post implementation workflows. This will allow us to compare the anticipated new workflows to the actual post implementation workflows. In future work we will also observe the pre-implementation workflows in the clinic to complement the stakeholder interviews. Conclusion: There is a need for a structured approach to mHealth platform implementation in clinical settings. Translating existing HIT workflow assessment tools to the mHealth domain has shown tremendous value in guiding our implementation process.

Using OpenEMR in HIT Training at Columbia

Virginia Lorenzi, MS, CPHIMS 1,2, Anthony McCormick3,4, Brady Miller, MD3, Kenneth Newman, DMD, PMP5, Carol Smyth, MB6; Fred Rahmanian, MS7, Frances Morrison, MD, MPH, MA2
1Department of Information Technology, NewYork-Presbyterian Hospital, New York, NY; 2Department of Biomedical Informatics, Columbia University, New York, NY; 3OEMR; 4Medical Information Integration, LLC; Catholic Health Services of Long Island; 6Medivo, Inc. New York, New York; 7KPMG, Inc., Malvern, PA

Introduction

The Columbia Certification of Professional Achievement in Health IT is a nine-month, credit-bearing program that provides broad training in Health IT to a professionally diverse student body. A goal of the program is to provide students with hands-on lab experience using current, relevant technology. From the initial funding in 2011 as part of the ONC Workforce University Based Training program, we looked at working with various vendor and free online EHRs, but these were expensive, difficult to access, or un-configurable. We decided to use OpenEMR, which was not only free and open source, but also ONC-certified. OpenEMR has provided a readily available, configurable platform that has added benefit to the students’ experience in many ways.

Methods

OpenEMR was installed on a server external to the closed University and Hospital systems in order to be accessible to students and external instructors. The system became an easily accessible playground that allowed our students to play many different roles, from user to system administrator. Students became accustomed to EHR functionality by entering fictional patient data, which was added to imported de-identified data. Later assignments included querying the OpenEMR database. Because OpenEMR was ONC 2011 certified, it was ideal for coursework that involved implementation for meaningful use; students created a mock project plan that involved an implementation in a fictional clinic.

Students also had the opportunity to collaborate with the open software community in real-time challenges. As the nation rushed towards ONC 2014 certification and Meaningful Use Stage 2 attestation, OpenEMR also struggled with the challenge of upgrading to an ONC 2014 system. This challenge turned into a real-world problem scenario for the class, who found that their EHR was not certified in time for their mock clinic’s go-live date. The students learned about the Meaningful Use Stage 2 implementation struggles and the hardship and flexibility regulations. This enabled students to don the vendor role; they joined OpenEMR conference calls and were given real assignments that helped in the development and certification of an ONC 2014 certified version of OpenEMR.

Conclusion

Open source software provides an accessible, quality playground for students. We have found that using OpenEMR has helped us solidify a multitude of skills and has provided some unintended benefits. Our students learned about how open source is used throughout Health IT within products and how vendors value-add services. They were able to explore the EHR user, implementer, and developer role. They were able to build relationships within the OpenEMR community and perform useful work that they could use on their resume.

We feel that we can expand the scope of assignments that use OpenEMR, including projects related to interoperability, patient engagement, form design, analytics, decision support, and security and privacy. Since OpenEMR is a real system used in healthcare, we anticipate that it will continue to provide opportunities for just-in-time real world activities to enhance our students’ learning experience.
Designing a Plan Do Study Act Framework to Promote Proper Utilization of Early Detection Technology in the Acute Care Setting

Graham Lowenthal¹, Patricia C. Dykes, PhD, RN¹, Stuart R. Lipsitz, ScD¹, Catherine S. Yoon, MS¹, Ronen Rozenblum, PhD, MPH¹, Perry G. An, MD², Suzanne Salvucci, RN², Christine Shaughnessy, RN², David W. Bates, MD, MSC¹,³

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

Introduction: Patient monitoring systems allow for early detection of clinical decompensation and have been associated with improved quality of care, lower length of stay, and favorable healthcare cost outcomes.¹,² In January of 2013, an early detection system was implemented on all adult medical-surgical units in an Eastern Massachusetts acute care hospital. The purpose of this study was to identify barriers to effective use of the early detection system and, using a Plan Do Study Act (PDSA) framework, develop continuous quality improvement strategies to promote proper utilization.

Methods: We established a committee of key stakeholders, including physicians and nurse champions from all study units, researchers, and vendor partners. We (the committee) conducted regular meetings to identify barriers to optimal system use. We designed a multi-cycled PDSA framework to address each barrier by identifying specific strategies and tactics. We tracked use of the system and response times (in minutes) to system alerts to measure progress. At each committee meeting, we reported out on progress and gathered feedback to evaluate the effectiveness of each strategy. We then adjusted our approach, set new goals, and executed revised tactics.

Results: We identified two major categories of barriers that could hinder effective system use. Educational barriers were those caused by a lack of awareness of the existence and importance of certain aspects of the system, and/or by unfamiliarity with using the system and features. Clinical Workflow barriers were related to challenges incorporating use of the system into everyday operations and communication.

We developed and tested several strategies for overcoming each barrier. To address educational barriers, we hosted a kickoff meeting to emphasize the importance of the system and proper utilization. We provided regular didactic and hands-on teaching sessions and coached staff on proper alerts management. In addition, we regularly reviewed recent clinical “saves” made using the device to demonstrate perceived system value. For clinical workflow barriers, we improved communication by integrating scripted discussion of the system, alerts, and trends into multidisciplinary rounds, EHR documentation, and handoff reporting. We mitigated alarm fatigue by enabling nurses to change device settings within specified limits to reduce unnecessary alerts and by providing several methods for receiving alert notifications (pagers, bedside monitors, and central nursing stations). We also strengthened our partnership with the vendor to address issues and concerns with the technology. We have noted improved response time trends to system alerts. Monthly overall mean response times have improved from 92 (range: 17 to 360) minutes at baseline to 54 (range: 4 to 127) minutes at present, mean Cardiac alert responses improved from 105 (range: 17 to 312) to 25 (range: 4 to 105) minutes, and mean Respiratory alert responses improved from 91 (range: 26 to 360) to 57 (range: 7 to 127) minutes.

Conclusions: Overall, our PDSA framework was effective for addressing barriers to proper utilization of early detection technology in the acute care setting. Since implementing the PDSA methodology, we have seen continuous improvement in system use and alert response times. We will continue using our PDSA framework to support proper system utilization and to decrease unit response times.

Acknowledgements: The authors would like to acknowledge the clinical champions at Newton Wellesley Hospital and the support for this project from EarlySense, Ltd.

References
Variability in the Sequence of HL7 2.x Event Code Types used to Represent Encounters Across a Health Information Exchange

Tina Lowry, MS\textsuperscript{1}, Eugene Kim, MD\textsuperscript{1}, Joseph Johnson III, MBA\textsuperscript{2}, Jason S Shapiro, MD\textsuperscript{1}  
\textsuperscript{1}The Department of Emergency Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA  
\textsuperscript{2}J2 Interactive, Charlestown, MA, USA

Introduction:  
While developing encounter-based measures (frequent emergency department (ED) visits, early (72-hour) ED returns and 30-day hospital readmissions) across a health information exchange (HIE), we discovered a high degree of variability in the sequence of Health Level 7 (HL7) 2.x event code types (A01, A02, etc.) used to represent different types of encounters. This posed challenges in implementing encounter-based measures in a standardized manner. The objective of this study was to analyze and describe the degree of variability in the way HL7 2.x messages are used to represent three general types of encounters across an HIE: 1) treat and release ED visits, 2) ED visits leading to an inpatient admission, and 3) other types of encounters (direct admissions, same day surgeries, etc.).

Methods:  
HL7 messages were randomly selected from 30 of 31 acute care sites in the Healthix HIE in New York from 12/29/2014 – 02/02/2015. One site that did not use standard pipe-delimited separators in their HL7 formatting was excluded. The data was parsed using the Python library HL7APY with regular expressions applied to search for message segments and fields, creating a data set of patient encounters with the following data elements: HL7 event code type, acute care site, visit number (unique encounter identifier), medical record number (MRN – a unique patient identifier), patient class (nominally designates emergency, inpatient or outpatient visit type), and message date and time. Data were imported into SQL Server for further analysis and grouped by facility, MRN and visit number. Each group was then sorted by message date and time and assigned a sequence number. Event types were compared to patient class to show variation and better classify the sequence of event types. Message sequences were then manually reviewed by 2 reviewers to determine which sequences corresponded with each of three general types of encounters described above.

Results:  
We extracted 3,363 sample HL7 messages representing 323 patient encounters from 30 sites. Manual review yielded 32 patterns of HL7 event code type sequences: 5 unique patterns for treat and release ED visits, 18 for ED visits resulting in inpatient admissions, and 9 for all other visits. Each pattern not only gave insight into different workflows for each site but also the frequency of use and interpretation of specific event type codes. For example, A04 (register a patient) is generally used to indicate patient registration in the ED and begin the patient encounter sequence. This event type was used by all 30 sites for emergency visits. An A03 (discharge/end visit) event type code is normally used to end a visit. However, when ED visits resulted in inpatient admissions, only 2 sites used A03s, while 16 used A06s (change an outpatient to inpatient) or A02s (transfer a patient), 1 site used A11s (cancel admit/visit notification), 7 sites used A08s (update patient information), and 4 sites use direct admit A01s (admit/visit notification).

Conclusion:  
The high degree of variability in the sequence of HL7 2.x messages used by acute care facilities to represent different types of encounters may confound attempts to apply encounter-based measures to an HIE dataset without further normalization. The sequences described in this analysis will be used to map all encounters in the HIE dataset to standard encounter types in order to validate and more accurately apply encounter-based measures. This issue likely exists in other HIEs where standard HL7 2.x has been implemented across multiple acute care facilities. In the future, use of XML-based standards (e.g., C-CDA) to aggregate encounter-based information in an HIE may obviate the need for this type of analysis and normalization, but further investigation will need to be done to determine if similar issues exist regardless of messaging standard employed.
Generating the MEDLINE N-Gram Set

Chris J. Lu, Ph.D.1, 2, Destinee Tormey1, Lynn McCreedy, Ph.D.1 and Allen C. Browne 1
1National Library of Medicine, Bethesda, MD; 2Medical Science & Computing, Inc, Rockville, MD

Abstract
The MEDLINE n-gram set is a very useful resource in Natural Language Processing (NLP) and Medical Language Processing (MLP). Currently, there is no MEDLINE n-gram set available in the public domain. Due to the large scale of data, it is a challenge to generate MEDLINE n-grams to fit into a research schedule with limited computer resources. The Lexical System Group (LSG) developed an algorithm to generate the MEDLINE n-gram set for adding multiwords into the SPECIALIST Lexicon. We believe the NLP community can benefit from access to this big data. We processed 2.6 billion single words from 22.4 million MEDLINE documents (titles and abstracts) to generate MEDLINE n-grams (n = 1 to 5) with terms appearing at least 30 times and having less than 50 characters for the 2014 release.

1. Background, Motivation, and Challenge
The LSG used the MEDLINE n-gram model to generate multiwords in 2014. A sophisticated algorithm has been developed to generate a comprehensive MEDLINE n-gram set to replace the approximated n-gram model by prediction filter [1]. N = 1 to 5 were chosen because these terms cover up to 99.47% of the terms in the Lexicon, as shown in Table-1. Also, 99.55% of terms in the Lexicon have less than 50 characters and thus terms with more than 50 characters are filtered out. Both word count (WC) and document count (DC) are calculated. The approach of deriving n-grams by bigrams doesn’t provide accurate WC/DC. Titles and abstracts from 22,356,869 MEDLINE documents are collected and then tokenized to 126,612,705 sentences and 2,610,209,406 words (tokens) in the 2014 release. Our sentence tokenizer found very few unrecognized sentence patterns (0.01% - 14,314). Java class (HashMap) is used for the key-value mechanism to store n-grams as keys and their associated information (WC, DC, and PMID) as values in the process. This system works perfectly for unigrams and bigrams. However, this approach does not work for n-grams where n > 3 because the number of terms (keys) exceeds the limit of maximum keys in the Java HashMap class (230-1). The processing time is not feasible (it can take months) when using persistent key-value DB or other database (embedded or server) approaches to replace the Java HashMap, even with a powerful computer that has 192 GB memory.

2. Approach – Split, Group, Filter, Combine, and Sort
An algorithm was developed to resolve the issues stated in the previous section. First, it splits the MEDLINE documents into S sections so that n-grams of each section can be generated within Java limitations. Note that duplicated terms are distributed across S sections. Second, it retrieves unique terms (keys) and calculates WC and DC by grouping terms from a range of alphabetic characters, such as terms from ‘c’ to ‘f’, through generated S sections. Each range of grouped (unique) terms, which are within Java limitation, is saved into G groups. Third, it applies WC and DC filters through all groups and then combines all filtered terms to form the n-gram set. Fourth, it sorts the n-gram set by DC, WC, and the alphabetic order of terms. For example, 20 sections of 5-grams were generated from MEDLINE titles and abstracts. Then, 14 groups with different ranges of alphabetic characters of unique 5-gram are retrieved from these 20 sections. The program combines the 5-grams by filtering out the low frequency terms from these 14 groups. Table-2 shows 1.54M 5-grams with WC >= 30 are retrieved from 1.67G 5-grams in this step. Finally, these filtered 5-grams are sorted and saved in a file. The idea is to keep those large numbers of invalid n-gram terms (N = 3 to 5) in different sections and groups, to resolve the Java limitation issue. The WC and DC filter is then applied to remove these low frequency (invalid) n-gram terms before the final combination.

3. Conclusion
The 2014 MEDLINE n-gram set is successfully generated by the above method. Table-2 shows the number of each n-gram. Three fields (document count, word count, and n-gram) are included. The MEDLINE n-gram set is distributed annually by the National Library of Medicine (NLM) with the SPECIALIST Lexicon annual release via an Open Source License agreement [2].

References
The Role of Technology Utilization in Designing Self-Management Systems

Robert J. Lucero, PhD, MPH, RN\textsuperscript{1,2}, Rosario Jaime-Lara, MS, RN\textsuperscript{3}, Yamnia Cortes, MS, MPH, RN\textsuperscript{4}, Dante Tipiani, MSW\textsuperscript{5}, Maribel R. Granja, MPH\textsuperscript{4}, Suzanne Bakken, PhD, RN\textsuperscript{4,6}  
\textsuperscript{1}College of Nursing, University of Florida, \textsuperscript{2}Center of Innovation on Disability and Rehabilitation Research, VA HSR&D, Gainesville, FL; \textsuperscript{3}School of Nursing, University of Pennsylvania, Philadelphia, PA; School of \textsuperscript{4}Nursing and \textsuperscript{5}Depts of Medicine and \textsuperscript{6}Biomedical Informatics, College of Physicians & Surgeons, Columbia University, New York, NY

Abstract

We report preliminary findings associated with the development of an electronic family-health information management system (eFHIMS). A directed content analysis of family caregiver focus groups was conducted to identify the extent to which Hispanic caregivers apply and perform self-management skills and tasks, respectively, as a basis for generating participatory design activities. Technology use was mentioned more than any other skill by caregivers to address and/or accomplish self-management tasks.

Introduction

Technology-based caregiver interventions can improve decision confidence, reduce emotional strain, improve spousal relationship conflict, decrease activity restriction, increase self-efficacy, and decrease burden\textsuperscript{1,2}. Nonetheless, electronic information management systems can assist caregivers manage health care services encounters and execute healthy behaviors for themselves and a family member living with dementia. The purpose of this poster is to present preliminary caregiver focus group findings associated with the development of an eFHIMS.

Methods

Based on a multi-phase system design approach, Consumer-centered Participatory Design, we conducted four focus groups (i.e., 2 English and 2 Spanish)\textsuperscript{3}. Audio recordings were transcribed and coded in the language of the group. A directed content analysis was conducted using the concepts present in Lorig and Holman’s self-management framework. The aim was to identify the extent to which Hispanic caregivers apply and perform self-management skills and tasks, respectively, in the context of health and health care as a basis for the participatory design phases.

Results

Caregivers were mostly female (90\%) with an average age of 81 years, more likely to be a daughter of a person living with dementia, and had been caregiving on average for 7 years. Caregivers expressed using or needing all of the self-management skills (i.e., problem solving, decision making, resource utilization, formation of patient-provider partnership, action planning, and self-tailoring) to accomplish the three self-management task (i.e., emotional, physical, and role management). Caregiver’s ability to utilize resources was the skill most often mentioned (n=73) to address self-management tasks. Notably, technology utilization, a sub-classification of resource utilization was mentioned more (n=43) than the other skills which ranged from 15 for self-tailoring to 33 for decision-making.

Conclusions

Hispanic caregiver’s technology utilization will be an important consideration as we develop the eFHIMS. This knowledge will be used to motivate design activities to elicit requirements of information technology use.

References


Acknowledgements: This study was funded by R01NR014430.
Bridging the Representation Gap of Medical Image and Clinical Note through Semantic Association Mining

Jake Luo1,2 PhD, Timothy Patrick1,2 PhD
1 Department of Health Informatics and Administration, University of Wisconsin Milwaukee, Milwaukee, WI; 2 Biomedical Data and Language Processing Center, University of Wisconsin Milwaukee, Milwaukee, WI

Abstract: Medical images and text are two of the most important forms for delivering clinical information and knowledge. Automatic integration of image knowledge and text knowledge can greatly assist clinicians in diagnosis, biologists in searching for evidence. However, the lack of a shared knowledge representation for images and text poses a significant barrier to the development of new applications that take advantage of this potential integration. This project aims to bridge the gap by systematically integrating image-text information to develop a knowledge base using natural language processing, medical image processing, and ontology-based knowledge integration methods. The image-text knowledge base can be a great help for clinical and translational research. For example, physicians can use the knowledge base to create semantic queries for searching patients with two criteria: one specified in clinical note “diagnosed with brain tumor in the past three months”, and the other specified in the image “nodules larger than 5mm at the pineal gland on MRI”.

Method: In stage 1 (See Figure 1 stage 1), we used the University of Washington Radiology case dataset to extract radiology report titles, figures, clinical notes, and captions. In stage 2, we parsed all the images and clinical notes from the datasets and store them in a database. We used a semantic annotator to automatically extract medical concepts from the clinical notes that described the medical image. The semantic relationships of the medical concepts were identified using a natural language processing method. The association of image components and textual clinical concepts were linked manually. In stage 3, we extracted semantic predicates from the annotation. For example in figure 2, we can also aggregate all patient cases that had radiosurgery on the brain through the reasoning of RDF triples “Therapeutic Procedure -> treats -> Neoplastic Process” and “Neoplastic Process -> Occurs_at -> Body Part”. We call these image-enriched semantic predicate. We stored all the predicates into a knowledge base using semantic web technologies (e.g. RDF store) for convenient information retrieval and analysis.

![Figure 1. System processing stages](image1)

![Figure 1: A parsed sub-tree](image2)
Assessment of impact of guideline changes in CDS for cervical cancer screening and surveillance

Kathy L. MacLaughlin, M.D.1, Ravikumar K.E., PhD,1 Kavishwar B. Wagholikar, MD, PhD, Marianne R. Scheitel1, Hongfang Liu, PhD1 and Rajeev Chaudhry, MBBS, M.P.H.1
1Mayo Clinic, Rochester, MN, USA 2Massachusetts General Hospital, Boston, USA

Introduction
Complexity in cervical cancer screening and management guidelines and lack of personalized care recommendation systems poses a significant challenge to appropriate individualized care recommendations in cervical cancer screening and surveillance care process models. Clinical decision support systems (CDSS) offer a great potential in mitigating some of the challenges thereby improving cervical cancer screening and surveillance guideline compliance. At Mayo Clinic, we have developed a CDSS to automate standard guidelines based on information available in electronic medical records (EMRs) for cervical cancer screening and surveillance at the point of care. However, revision of these guidelines has posed a significant challenge to the development of the CDSS. In this work, we studied accuracy of the CDSS system after updating it to implement the latest revision of the guidelines. In an earlier study1, we assessed the performance of a CDSS for cervical cancer screening and surveillance.

Methods
The system extracts 10 data elements such as age and history of CIN 2/3, test values from both structured resources such as problem list and unstructured data sources such as cervical cytology report and clinical notes. It then makes care recommendations based on a set of decision rules2 (more than 55 decision scenarios). The rules were implemented in Drools rule engine3. We extended our earlier implementation1 by adding additional decision rules.

Data set and Validation
The CDSS system was run on 2,521 patients. Not all of the possible end points were represented in the batch sample that was acquired. We performed a stratified sampling on 191 of those patients, covering every decision end-point based on the extracted data elements (cytology, high-risk human papilloma virus and colposcopy biopsy results). The clinician expert evaluated the CDSS-generated recommendations by reviewing the medical record for each unique patient and compared them against ASCCP guidelines4, which allowed identification of errors in the CDSS.

Results and Discussion
Of the 191 patients reviewed, the CDSS-generated decision was correct in 151 cases, leading to an overall accuracy of 82%. Figure 2 outlines the error percentage distribution as per the decision endpoint. Analysis of errors revealed that in four patients, even though the final recommendation was correct, the process flow and endpoint was incorrect. We observed a drop in the accuracy by 5% when compared to earlier reported results by our group (accuracy of 87%). The drop in the performance of CDSS may be due to the increased complexity of the updated ASCCP guidelines4.

Conclusion
In this study, we evaluated the performance of the updated CDSS. While the revision of the practice guidelines resulted in increased complexity that impacted the performance of the system, a systematic error analysis coupled with stratified sampling helped us understand the micro behavior of the CDSS. We soon plan to deploy the system in big data environment and perform a large-scale evaluation on the impact of clinical care recommendations for cervical cancer prevention at the point of care.

Acknowledgement: Agency of Health Care Research and Quality (AHRQ R21HS022911) and National Library of Medicine, NIH (K99LM011575) supported this research.

References
A Framework for Incorporating Changes to a Reference Terminology on a Mapped Enterprise Terminology Subset

Sina Madani, MD PhD\textsuperscript{1}, S. Trent Rosenbloom MD MPH\textsuperscript{2}
\textsuperscript{1}The University of Texas MD Anderson Cancer Center, Houston, TX; \textsuperscript{2}Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Introduction
Enterprise vocabularies that are mapped to or subsetted from reference terminologies can be used within health care organizations for documenting and representing common patient data at the point of care. Such custom vocabularies are often developed internally, enriched with domain-specific terms, and linked to concepts from external reference terminologies. However, reference terminologies are updated frequently by their publishers. This can contribute to the burden of managing updated codes to the health care organizations.

A patient problem knowledge base (PKB) vocabulary subset of SNOMED CT has been in use at the Vanderbilt University Medical Center since 2011. All 112,319 concepts in this terminology have been mapped to equivalent SNOMED CT concepts, locally augmented with additional terms extracted from patient notes, and maintained in a commercial terminology server. Periodic synchronization of PKB terminology mappings to the latest release of SNOMED CT is essential for ongoing Vanderbilt EMR system and operational use.

Methods
From 112,319 PKB concepts, all concepts with retired SNOMED CT 2013 codes (2,692) were retrieved from the terminology server. For these retired PKB concepts, we identified five SNOMED CT historical relationship types in our terminology server and retrieved the mapped concepts. We acknowledged the updated SNOMED CT (2014) target concept codes for the retired PKB concepts from 4 relationships and only reviewed 300 PKB concepts with the retired SNOMED concept that had a “May be a” historical relationship (Table 1). The two authors reviewed the updated concept codes and associated terms for compatibility with the exiting PKB terms, using a four value voting system. Disagreements were reconciled iteratively through discussion, with persistent disagreements reported here.

Results
Based on the updated version of SNOMED CT, 387 target terms from “May be a” relationship were identified for the 300 PKB concepts with retired SNOMED CT codes. We compared these targeted new terms with the existing 2,764 PKB terms associated with 300 concepts and evaluated their correctness (Table 2). The disagreement rate on the suggested new terms was decreased from 45 concepts to 32 and 9 respectively after two reconciliation sessions between the two reviewers. “Miscarriage”, “Termination of Pregnancy”, and “Pregnancy with abortive outcome” were the main sources of disagreements among representational variations of the 9 concepts with 22 terms.

<table>
<thead>
<tr>
<th>SNOMED relationship</th>
<th>concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was a</td>
<td>1845</td>
</tr>
<tr>
<td>May be a</td>
<td>300</td>
</tr>
<tr>
<td>Moved to</td>
<td>291</td>
</tr>
<tr>
<td>Same as</td>
<td>185</td>
</tr>
<tr>
<td>Replaced by</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 1. SNOMED retired concepts relationship

<table>
<thead>
<tr>
<th>Ruling on targeted terms</th>
<th>Agreement</th>
<th>Disagreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>93</td>
<td>3</td>
</tr>
<tr>
<td>More general</td>
<td>83</td>
<td>9</td>
</tr>
<tr>
<td>More detailed</td>
<td>168</td>
<td>4</td>
</tr>
<tr>
<td>Questionable</td>
<td>32</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>376</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 2. Verdict results on target SNOMED terms against PKB terms

Discussion
We will use the results of this study for the proper update to the mappings of our internal vocabulary system. Suggested new mappings with questionable outcome will be re-examined against the future versions of SNOMED CT and updated accordingly. Although the terms in questions were reasonably consistent with their definition in the Unified Medical Language System, the ambiguity in their representation made it difficult for the reviewers to judge universally. This overlaps with the issue of synonym drift that has been described elsewhere in the terminology mapping domain.
Title
Choosing to Build: Optimizing the Development of a Custom Pathology Laboratory Software Solution

Authors
Emilio Madrigal¹, Victor Brodsky²
¹Department of Pathology, Mount Sinai Health System, St. Luke’s, Roosevelt, and Beth Israel Hospitals, New York, NY, USA
²New York-Presbyterian Hospital/Weill Cornell Medical College, Pathology, New York, NY, USA

Background
Commercially available software seldom meets all the requirements of laboratory medicine. Customizations by vendors can result in additional costs and compromises, while using software as-is may lead to inefficient workflows and impact patient safety. At NYP/Weill Cornell we chose to create a responsive, web-based, multi-user application for submitting and tracking scan orders to a new whole-slide scanning service. We analyzed the amount of time and effort to design, build, and test a working solution, focusing on how prioritization of issues documented during development impact the time to completion in days (TC).

Design
Bitbucket was used as the distributed revision control system and issue tracker. Source code statistics for the project, built on Microsoft Windows 2012 with PHP, were generated using GitStatX. Developer productivity was measured using lines of code (LOC) written per-day by our software engineer, who holds a Master of Biomedical and Computer Engineering. A physician-informaticist (PI) responsible for conceptual design, functional specifications, and program testing was surveyed regarding his time commitment.

Results
The development team comprised 1.2 FTEs; the PI dedicated 0.2 of his time, spending 60% of it on software testing. The project took 451 days and contains 1999 files with 379,620 LOC (averaging 670/week). In total, 377 issues were documented, of which 325 were resolved, and 49 remained outstanding at launch; 204 (54%) were classified as bug, 163 (43%) as enhancement, 3 (1%) as proposal, and 7 (2%) as task. The top two priority flags assigned to bugs and enhancements were “blocker” and “critical”. The average TC for all issues was 29 days (13 for bugs, and 52 for enhancements) with a median of 9, and SD of 59.2. “Blocker” was used 24 times: 19 (79%) for bugs with a TC of 3.4, and 5 (21%) for enhancements with a TC of 6.6. “Critical” was used 221 times: 139 (62.9%) for bugs with a TC of 12, and 78 (35.3%) for enhancements with a TC of 39.2.

Conclusion
Although testing and issue documentation required a generous amount of time, prioritization positively impacted the rate at which development milestones were achieved. Issues marked with the highest priority of “blocker” were resolved 3 to 6 times faster than those flagged as “critical”. Additionally, daily LOC logs established a baseline expectation of activity. The use of issue tracking and version control systems provides an effective way of monitoring development in order to efficiently build software solutions.
Quantitative Evaluation of Dysarthria and Development of Vowel Sound Voice Training System

Hiroyuki. MAEDA, BS¹, Naoya. ARISAKA, MS², Wakana. HATA, MS², Noritaka. MAMORITA, PH.D¹, Ikuyo. ISHIZAKA, PH.D², Kazuhiko. YAMASHITA, PH.D³, Harukazu. TSURUTA, PH.D¹

¹Kitasato University Graduate School of Medical Sciences, Sagamihara, Kanagawa, Japan; ²Kitasato University School of Allied Health Sciences, Sagamihara, Kanagawa, Japan; ³Tokyo Healthcare University Division of Healthcare Informatics, Setagaya, Tokyo, Japan;

Introduction

Dysarthria is a disorder of voice and articulation. It occurs with high probability after cerebrovascular disease, which eventually lowers patients’ quality of life. Therefore, rehabilitation and its evaluation are very essential for the patients’ quality of life after discharge. However, quantitative assessment of the dysarthria rehabilitation is not widespread because the instruments used for measurement at hospitals are expensive and not portable. To solve this problem, we developed a voice training system which works on usual PCs.

Methods

We developed an application which can record the patients' voices during vowel sound voice training, measure the duration and volume of the recorded voice, and calculate the formants based on frequency analysis. This application also has a training feature which provides visual feedback of the patients input to encourage the patient. The program was developed using Visual Basic 2010. The evaluation study was approved by the Ethic Committee of Kitasato University Hospital and all participants provided their informed consent for participation.

Results

Figure 1 shows an example of the output of the vowel sound voice training, in which a subject with cerebral infarction (hoarseness, non-hoarseness) was able to continue the training for 2 or 3 seconds, and it was found that the power spectrum of the second formant frequency band with non-hoarseness is larger than hoarseness. Visual output was able to be used as a bio-feedback interface, which encouraged the patient to utter louder and longer.

Conclusion

We developed an application which can be used for the quantitative assessment of dysarthria and the vowel sound voice training. To evaluate the usefulness of the system, we are studying cerebral infarction patients.

References

Designing a drawing-based tool to manage EBRT process in an open-source oncology EMR system

Manika Maheshwari, Saptarshi Purkayastha
School of Informatics & Computing, Indiana University-Purdue University, Indianapolis, USA

Problem
External beam radiotherapy (EBRT) is a complex treatment process that requires appropriate communication and planning between the oncologist, interventional radiologist and the radiology technician. Hand drawn images or annotations on radiology images are often used for communication. Cone beams, angles, treatment ports, dose distribution curves are in a state of constant flux particularly when used in conjunction with modern image-guided (CT or X-ray) radiation therapy. Treatment procedure needs to be well articulated and communicated by electronic means and remains a challenge in general-purpose EMR systems.

Purpose
This poster showcases our community-based participatory research to implement an EMR system to manage care provided in oncology clinics by integrating the workflow of radiation practices. The goal of our research is to explore the challenges and opportunities of using open-source tools to support an oncology clinic in a large teaching hospital, particularly in their radiation therapy process. Based on the user needs, we designed and developed a drawing module that would facilitate better communication and reduce errors.

Methods
We implemented the oncology terminology in the OpenMRS, the open-source EMR system. The global community of informaticians participated in customizing a dictionary that was suitable for oncology practices. We then involved interventional radiologists to further customize to add granular terms that are used in their practice. We observed the practice of hand-drawings, and the use of markers on X-rays that is used by radiology oncologists and technicians to communicate.

The drawing module was created by a Google summer of code intern to allow adding images and add pen-based input on these images. The images came from a PACS or RIS system into the EHR, with ability to make annotations and drawings on the images. These CT/X-ray images are also updated in a real-time basis on which the oncologist or radiologist can make changes that are reflected on the technician’s screen. The EMR system can also generate dose distribution curve based on a set of values that are entered by the oncologist. Over the next few weeks of treatment, based on the images that are recorded in the PACS, we also implemented a feature where dose distribution curves that can be adjusted by the oncologist in the drawing module and gets reflected on the technicians screen.

Preliminary results
Opportunities – The open-source community allowed the development of a quick terminology dictionary for oncology practice. The network of Informaticians provided suggestions for the creation of forms to generate the dose distribution curves. The development of the drawing module was done by the support of Google for open-source projects. Students participated in the community and developed tools that can be used by any implementation around the world.

Challenges – The limited capabilities of integrating OpenMRS with EBRT devices hampers direct management. Technicians are skilled users of EBRT devices, but not experienced in the use of EMRs. Device integration remains a challenge unless access to device APIs is available to developers of EMR system.
User-Centered Design of an Application to Aid in the Safe Return to Work of Injured Farm Workers

Andrea Mahnke, MS, Laurel Verhagen, BS, Bryan Weichelt, MS, MBA, PMP, Iris Anne Reyes, MPH, Will Ray, BA, Matthew Keifer, MD, MPH

(1)National Farm Medicine Center, Marshfield Clinic Research Foundation, 1000 N Oak Ave, Marshfield, WI 54449
(2)Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, 1000 N Oak Ave, Marshfield, WI 54449

Background
Large animal farms are growing in size, increasing work task specialization and immigrant labor. Workers not only face inherent risks in the agricultural workplace, but are also introduced to significant dangers in these operations. Injuries in dairy and pork farms are common and are increasingly managed by primary care physicians. Yet, clinicians are often unfamiliar with the physical demands of farming and have little training and few resources to manage the safe return to work of injured workers. This project will develop a computer application designed for clinicians, working with patients, to guide safe return to work planning for injured workers in the dairy and pork industries.

Objective
We employed user-centered design methods to build an application to assist clinicians, physical therapists and occupational health therapists in translating physical restrictions into potential work task recommendations specific to an injured worker. The application has the potential to replace existing ink over forms and would result in more structured data, which helps make the data available for research projects. Due to the large population of Spanish speaking workers, the application’s interfaces and forms will be available in both English and Spanish. The flow of the new application starts with an industry, agriculture or other, entering job task physical demands into the system. Then when an injured worker is being assessed physical limitations will be entered by the clinician. Next the system will generate a list of potential tasks the injured worker may be able to do. The clinician and injured worker would then select tasks that are a match for their job. An output sheet would then be generated which would include physical limitations, as well as recommended tasks for the injured worker to take back to employer. The injured worker and employer would review the list and arrive at appropriate tasks for the worker to do while healing from injury.

Methods
This project employed an iterative user-centered design method on three primary end user types that interact with the system. These included employers, workers and the clinicians, including physical and occupational therapists. A usability analyst visited dairy farms to interview farmers regarding their experience with returning injured farm workers to work. Focus groups were also conducted with farm workers, some of whom had been through the injury and return to work process. Audio recordings were taken for all sessions; focus groups were also video recorded. All participants were compensated for their participation. Interviews and focus groups probed the understanding and interpretation of the existing Workers’ Compensation Report that is given to both farmer and worker by the clinician. The current form identifies physical restrictions for the worker, but does not make any recommendations regarding suggested tasks the worker may be able to perform while healing from injury. Interviews were also conducted with a variety of clinician roles that had experience with injured farm workers. In parallel with the interviews and focus groups, the team developed several iterations of a structured version of the form and a “patient friendly” output report. Patient friendly forms had a cleaner, less dense layout and included recommended work tasks. For both interviews and focus groups the participants were presented with the current form and a redesigned iteration of both the structured form, as well as a patient friendly form. Feedback regarding the alternate layout of the forms was collected. Qualitative analysis of interviews and focus groups was done using NVivo software. Areas for improvement were identified and implemented iteratively.

Results & Conclusions
Data collection and analysis ongoing. Valuable insights were revealed by focus groups and interviews that have helped to steer interface design of early application iterations.
The Process of Using Focus Groups to Inform Development
Jennifer A. Mallow PhD, FNP-BC1*, Laurie A. Theeke, PhD, FNP-BC, GCNS-BC1 Elliott Theeke1 and Brian K. Mallow, PMP, MCDBA2
1West Virginia University, School of Nursing 9620 HSC South, Morgantown, WV 26506
2 Sovern Run, LLC, Information Technology Consultant 210 Cardinal Lane, Albright,

Introduction

Used as an integrated tool, technology may improve the ability of healthcare providers to improve access and outcomes of primary care in rural populations (1). Incorporating input from the beginning stage of development to implementation of the intervention has the potential to enhance adoptability and sustainability in real-world settings. This poster reports the findings from focus groups aimed at evaluating a newly developed mHealth architecture called mI SMART (Mobile Improvement of Self-Management Ability through Rural Technology). The mI SMART platform combines multiple technologies including HIPAA compliant, web-based system of mHealth sensors and mobile devices that allows patients living in rural areas to perform self-monitoring, receive feedback in real time, attend eVisits for education, mental, and physical health from their home, and displays a record database to patients and providers. The creation of mI SMART is being guided by the model for developing complex nursing interventions (2).

Methods

In 2014 we completed focus groups in an academic, outpatient, and community setting. The approach for this study was developed using Krueger’s Focus groups: A practical guide for applied research (3). The focus groups were 60 minutes long. The facilitator provided a demonstration of the mI SMART platform. Directly following the demonstration, the following open ended questions were asked to each group: What do you think your biggest obstacle is when thinking about prescribing mHealth tools? As a patient yourself, what do you think your biggest obstacle is when thinking about using mHealth tools? What did you like best about the mI SMART platform? What would you like to see removed from the mI SMART platform? What would you like to see added to the mI SMART platform? Which outcome measures do you think would be most helpful to measure when treating patients with chronic illness? What outcome measures do you think are most important to patients? What else would you like to tell me about mI SMART? Lastly, the participants were asked to complete a questionnaire with demographic information and specific questions regarding the past, current and future use of mHealth tools. The field notes were entered into a Microsoft access database with the demographic information and responses to specific questions regarding the past, current and future use of mHealth tools. Several procedures were employed to maximize the transcription quality, and to ensure that quality standards were maintained (4).

Results

Both prescribing (N = 8) and non-prescribing healthcare team members (n=21) attended the focus groups. The majority were practicing for more than 20 years (44.8%) in an outpatient clinic (62%) for 20-40 hours per week (37.9%). Only 2 participants had used mHealth tools in the past. Providers identified perceived obstacles of patient use as ability, willingness, and time. System obstacles were identified as lack of integration, lack of reimbursement, and cost. Nineteen providers (65%) reported that they would definitely use this mHealth architecture in the future. The positive attributes of the developed system were: capability for virtual visits, readability, Bluetooth connectivity, user-friendliness, ability to capture biophysical measures, enhanced patient access, and incorporation of multiple technologies. Providers suggested increasing capability for biophysical and symptom monitoring.

Conclusion

Technology interventions have the potential to improve access and outcomes but will not be successful without the input of users. Findings from these focus groups are being used to inform continued development of the mHealth architecture prior our implementation of the system.

References

Responders and Nonresponders: Nurse Practitioner and Physical Therapist Personal Perceptions of Activity Monitors for Patient Use

Peggy J. Mancuso, Ph.D., RN, CNM, CNE, Sahiti Myneni, Ph.D., MSE
University of Texas School of Biomedical Informatics at Houston, Houston, Texas

Abstract
The purpose of this mixed-methods study was to explore personal perceptions of Nurse Practitioners and Physical Therapists regarding use of activity monitors as adjuncts to patient care. Reactions from a convenience sample of 23 participants pursuing doctorate education varied. Nineteen were positive for 14 weeks, but four became “nonresponders” at three weeks, a point when support may be needed for successful use. Most participants responded that these devices would be beneficial for some patient populations.

Introduction
Activity monitors, with associated smartphone applications and web pages, enable individuals to track activity, sleep patterns, and energy expenditure daily. The purpose of this study was to explore personal experience perceptions of Nurse Practitioners (NP) and Physical Therapists (PT) regarding the clinical utility, usability, and usefulness of an activity monitor (Fitbit® One (clip) and Force (bracelet) for patient populations. Such an understanding will allow utilization of these devices as effective components that form the core of patient-engaging care delivery processes.

Methodology
The following research questions were addressed: What are NP and PT personal experiences with daily activity monitors regarding utility, usability, and potential usefulness in patient populations? How do these perceptions change over time and with consideration of the existing evidence? A convenience sample of 9 NPs and 14 PTs students was accessed from practitioner students pursuing practice doctorate education. Each participant received a Fitbit® One or Force at the beginning of the semester. Participants were assessed weekly via voluntary discussion boards regarding ease of use and potential for devices to improve or worsen patient outcomes. Participants were requested to forward weekly activity reports provided by Fitbit® to investigators. At the end of the study period, participants responded to online surveys about the value of health informatics for patient care. One device used in the study, the Fitbit® Force was recalled during the study. This product recall enriched participants’ qualitative responses. Triangulation, rich description, and peer review ensured rigor of analyses.

Evaluation Results
There were 19 active NP and PT participants and 4 inactive NP participants. Inactive participants become “nonresponders” with minimal postings at approximately three weeks after initiation. Several themes emerged from qualitative analysis (see Figure 1): (a) usage timeline, (b) self-realization, (c) device relationship, and (d) potential use in patient populations. Five stages characterized the usage timeline: new user, technological difficulty, honeymoon period, habit formation, and habituated user. Those who expressed enthusiasm tended to use more device features (e.g., recording sleep) over time. Multiple users noted self-realization and surprise at activity levels. For example, “busy was not the same as being active.” Active participants described the device as a “motivational partner” that provided “validation.” Participants also described “doubts about accuracy” and “lack of trust.” They were frustrated with “lack of (technical) support” and “expense of the device.” Some participants expressed depression about the recall of the Fitbit® Force (e.g. “I miss my Force terribly”). The providers’ perceptions of potential use of these devices in patient populations varied by professions, with PT practitioners recommending generalized patient use and NPs proposing use for specific populations only. Both professions noted that activity monitors would be most successful with “motivated patients.”

Conclusions
This study explored current self-help technologies (activity monitors) from the perspectives of practicing PTs and NPs who were pursuing practice doctorates. Several socio-technical factors are found to play a role in the adoption of these technologies. Of interest, was the three-week period that marked the time that some NPs became “nonresponders” and ceased voluntary personal use of the monitor. This period could represent an important point when additional support would facilitate continued future use of the activity monitor.

References
ICan’tCount: a mobile app for helping children with dyscalculia
Andrea Marchesini¹, Vincenzo Maurelli¹, Siria Pantano¹, Giuseppe Vacante¹, Marilena Coniglio², Gianluca Palermo¹, Vittorio Tauro¹, Agostino Accardo³, and Sara Marceglia, PhD²,³
¹ Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano; ² Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy; ³ Dipartimento di Ingegneria e Architettura, Università degli Studi di Trieste, Trieste, Italy;

Introduction
Dyscalculia is a specific learning disorder that affects 1.2% of 7.1 million pupils in Italy and is usually recognized by the teachers who send the potentially affected child to a specialized neuropsychologist for diagnosis. Dyscalculia diagnosis relies on a paper-based test administered by the neuropsychologist. However, the paper-based test requires medical supervision, thus leading anxiety in the child. To avoid this, we designed and developed an app to support both the neuropsychologist and the child during the screening test and, after that, during rehabilitation.

Methods
Working together with the Pediatric Neuropsychology Unit of the Policlinico Hospital in Milan, we designed and developed an iOS application for iPad based on the AC-MT test, released by Erickson publishing¹. The App was designed to allow the clinician to see the results of the test and to assign specific exercises, in order to support parent’s rehabilitation. Parents were also considered as users and were allowed to both receive information on the disease and monitor the improvements of their child. The App was tested by three independent neuropsychologists during their normal ambulatory activity.

Results
The App, named I Can’t Count, includes five sections: ‘Gioca’ (Play) is the section in which the child completes the test; ‘Esercizi’ (Exercises) allows improving calculation skills through exercises assigned by the clinician, ‘Risultati’ (Results) shows the results of the rehabilitation activity, and in ‘Area Riservata’ (Doctor's Reserved Area) the clinician assigns the exercises to the child. The App, developed with an Italian graphic designer, displays virtual pages to make calculations as on a paper sheet. The three neuropsychologists used the App for one week in their clinical practice and reported that it helped the diagnostic process, supporting the children to easily and smoothly complete the AC-MT test. The children had no problems in understanding how to perform the test, thus fastening the screening visits.

Figure 1. Representative snapshots of iCan’tCount.

Conclusions
Even though still in a prototype version, our application could be a starting point for improving the diagnosis and rehabilitation of dyscalculia. The app was designed to have a user-friendly interface and to be used by doctor, children and their families.

References
Qualitative Analysis of Responses to a Questionnaire via an EHR Patient Portal

Elizabeth A Marshall, MD, MBA, Suparna Qanungo, PhD, Jihad S Obeid, MD, James C Oates MD, Melissa L Habrat, BA, Bob Warren, MD, PhD, MPH, Leslie A Lenert, MD, MS
Medical University of South Carolina

Abstract: Obtaining and updating patient preferences for participation in research can be cumbersome for both patients and clinical investigators. A patient portal to the electronic health record (EHR) provides a suitable venue for patients to indicate and manage their preferences. However the process may be confusing at times to patients. As a result unplanned questions and concerns may arise. We report on the qualitative analysis of patient responses thus far.

Problem addressed: Low recruitment into clinical trials historically has been a problem, resulting in incomplete study recruitment and/or increased costs to study sites. Increasing the number of volunteers for potential participation in research helps address this problem. Patients seen at a hospital or clinic typically have no means of indicating their interest in research participation. However it is the responsibility of health organizations to collect the permissions from patients who wish to be contacted for participation in future research studies, especially if they can be matched to EHR phenotypes. The Medical University of South Carolina (MUSC) has historically recorded patient research preferences via a paper consent form completed at the time of registration. Due to issues such as transcription error and lack of accessibility to update research preferences, a new method was developed.

Informatics intervention: This solution involves the use of a questionnaire sent via the EHR patient portal (Epic MyChart). MyChart had only been used for clinical correspondence within the hospital system. This was the first attempt involving research. During the implementation of this new method for research preference collection, a new project evolved involving the qualitative analysis of the replied responses to the questionnaire.

Methodology and evaluation: A questionnaire with various parameters to address participant’s research preferences were designed and distributed via MyChart. Data was collected from the onset of the first message received to the MyChart questionnaire between December 15, 2014- February 15, 2015. Interestingly, a total of 1108 e-mail messages were received via MyChart. These emails were unplanned responses to the questionnaire. They could be categorized into two groups: first, those who completed the questionnaire and sent an e-mail response; and second, those who just sent an e-mail response with no questionnaire. In order to do a qualitative analysis of all the messages/responses, a random sampling will be done such that 25% of the 1108 messages will be selected in total. Since messages came in at various sections of the day, an even sampling of responses from three time frames of Morning (4am-noon), Afternoon (noon-8pm), Evening (8pm-4am) will be taken. A traditional content analysis will be done comparing responses of those individuals that completed the questionnaire and those that did not; we will focus on the areas of comprehension, access to care, computer/technology, and insurance. With this approach, our goal is to capture all the diverse perspectives of the patients as well as identify barriers and concerns of using a patient portal to indicate and manage research preferences. The primary outcome of this study is to assess patient responses towards a research questionnaire in order to improve MyChart operations.

Conclusion: Patient-centric care approaches leveraging remote monitoring and informatics support represent a promising approach to maintaining patient satisfaction and clinical quality while reducing overall care costs. We anticipate that this study will contribute to the further evaluation of addressing some of the barriers of the use of a patient portal such as MyChart for future resolution of this increasingly important approach to improve recruitment into clinical trials.

Acknowledgments: This work was supported in part by the South Carolina Clinical & Translational Research (SCTR) Institute, with an academic home at the Medical University of South Carolina, NIH/NCATS Grant number TL1 TR000061.
Bring Your Own Device: From PDA To Smart Mobile Devices
Karen Martinez, BSN, MSc; Karen L. Courtney, PhD, RN; Elizabeth Borycki, PhD, RN & Noreen Frisch, PhD, RN, FAAN,
University of Victoria, Victoria, British Columbia, Canada

Introduction
The mobile nature of healthcare made the concept of mobile technology use attractive in the clinical setting. Personal mobile devices use in the clinical setting generates perceived concerns and benefits. Over the last thirty years, the types and capabilities of mobile technologies have changed significantly; however some benefits and concerns have remained constant despite the changes in technology. This poster will present results of a literature search done to describe and compare the use of electronic personal devices in the healthcare environment from the historical Personal Digital Assistant (PDA) to the modern smart mobile devices (smartphones/tablets). The poster will also include the similarities and differences in perceived concerns and benefits over time and extrapolate lessons that may apply to the next generation of mobile technologies. Next generation mobile technologies will continue to move forward. There is a need to address the benefits and risk of the use of these personal technologies as it can have an impact on the practice of healthcare professionals and patient care.

Methods
A literature search was performed (2000–2015) using PubMed, Google scholar, and Google. Grey literature (Google) was included as we were interested in subjective perceptions of benefits and concerns and empirical literature was limited when the devices were first introduced into the healthcare environment. Specific keywords for mobile devices included: “Personal Digital Assistant devices” or “PDA” as used in a “clinical setting” or “hospital” (2000-2008) and “BYOD [Bring Your Own Device]” or “tablet/iPad” or “smartphones” in “clinical setting” or “hospital” (2008-2015).

Results
The principles behind the past use of PDA devices and the current use of smart handheld devices, when compared, are quite similar. PDAs and smart mobile devices provide real-time access to information by providing decision support, education, and data collection at the point-of-care. PDAs enabled healthcare professionals to organize their contacts, calendars, but also access journals electronically. Unique to smart mobile devices was their use as a replacement for multiple devices health professionals previously carried.

Both types of devices were recognized for the benefits of convenience, mobility, and efficiency. Consistent concerns with both devices were privacy and security. Further, there was also a perceived issue with productivity that varied with the different types of devices over time.

Conclusions
The consumerization of these powerful handheld computerized devices has resulted in their affordability and the use of these personal mobile devices in the workplace is visible to many. The core benefits and concerns noted above seem stable across the different technologies. These factors are likely applicable to future personal technologies that may be brought into the healthcare environment by healthcare professionals. Healthcare organizations need to address integration of such devices to meet practice needs safely.
Simultaneous optic disc and macula detection using template matching

Maria A. Martínez-Castellanos, MD1, M. Elena Martínez-Pérez2, Tzolkin Garduño-Alvarado2, Luvia Rodríguez-Quíñones3, Samantha M. Salinas-Longoria1

1Asociación para Evitar la Ceguera en México I.A.P. Hospital "Dr. Luis Sánchez Buldnes", México; 2Institute of Research in Applied Mathematics and Systems, UNAM, México.

Abstract- Automatic computerized segmentation of optic disc (OD) and macula (MC) is an everyday need for clinical applications. Our algorithm achieves such a task by means of cross correlation of a template with the image. It performs a k-fold cross validation loop. Each of its iterations is composed of a training and a validating stages. Results are reported on 37 fundus images obtained from a diabetic retinopathy screening program.

I. INTRODUCTION

Diabetic Retinopathy (DR) is a chronic disease that affects the eye. More specifically, it is the manifestation of diabetes mellitus in the eye retina. Fundus images are used by ophthalmologists to trace and locate lesions caused by DR. The OD is a retinal area mildly oval with the major axis in the vertical position. Its mean area is 2.7 mm2 and its horizontal radius of approximately 1.8 mm. A variety of methods for the location and segmentation of the OD have been implemented, [1], [2], [3], to mention some. The MC is an oval area in the center of the retina. It is located 3 to 4 mm (15 degrees) from the OD, slightly downward, and between the major temporal vascular arcades. Some of the methods used for MC location can be found in [4], [5], [6].

II. MATERIALS

The set of images used in this paper consists of 37 images corresponding to either left or right eyes from patients suffering from diabetic retinopathy, images being captured with a Carl Zeiss FF450 plus IR fundus camera, the field of view angle used is 50 degrees and the size of the images is 489*647 pixels. They have had the OD and MC structures segmented by two licensed ophthalmologists.

III. ALGORITHM

Before making the computation of the process, a preprocessing stage is performed. Afterward, a kfold cross validation is executed on the images. A loop of k iterations is carried out, in each iteration k-1 of the subsets are used for training, while the remaining one is used for validation. Before running the algorithm, a noise reduction median filter is applied to the green band of the images, kernel size being 5*5. A polynomial approximation of the background illumination was made using a basis of discrete orthogonal polynomials, [7], [8], the polynomials used for the present approximation were of third degree. Next, the approximation was subtracted from the original image.

The training step of each iteration receives images, output being a grayscale template of the region where the OD and MC are located along with its corresponding coordinates (x,y) and the template where it belongs. The coordinates represent the center of a rectangular region the size where the OD and MC are located in the image.

IV. RESULTS AND CONCLUSION

The evaluation of the algorithm is a process with which the optimal k for the fold is chosen. Optimal k is chosen as the one that yields the least error. The error in this case is the distance between manual segmentation and the result of the algorithm. Two databases where calculated, a private one consisting of 37 images and the DRIVE dataset[9] consisting of 40 images out of which 35 were used given 5 did not contain OD and MC at the same time. On both datasets optimal k=3 was obtained. Our results showed 100% optic disc centers located within the OD area and 90% macula centers located within the MC area. The results in the DRIVE dataset showed 97% of the OD centers located within the OD area and 77% of MC centers located within the MC area.

ACKNOWLEDGEMENT

This work was financially supported by PAPIIT project number IN103414 from UNAM, Mexico.

REFERENCES

Abstract: This poster describes the study design for an innovative, pragmatic randomized controlled trial to evaluate the impact of a personalized, web-based patient portal, Inpatient myNYP. The federally funded trial is being conducted with cardiology patients at NewYork-Presbyterian Hospital/Columbia University Medical Center. The purpose of the intervention is to improve engagement and satisfaction in hospital patients.

Introduction: Patients who are better informed and more engaged in their healthcare have higher satisfaction and better health outcomes. While patient engagement has been a focus in the outpatient setting, strategies to engage patients while they are inpatient have been lacking. Patient-centered care can be improved in the hospital visit through better information management practices. We developed Inpatient myNYP, a personalized, dynamic web-based patient portal where patients can view members of their care teams, track current and historical vital signs, review their documented hospital medications and allergies, document pain scores, record personal notes, and send questions and comments to their care team members (Figure 1).

Methods/Results: We are conducting a pragmatic randomized controlled trial to evaluate the impact on patient engagement and satisfaction of giving patients access to Inpatient myNYP via tablet computers (Apple iPads) in 426 inpatient cardiology patients at NewYork-Presbyterian Hospital. Participants are randomized to one of three groups: 1) control group, 2) iPad with basic functionality, and 3) iPad with Inpatient myNYP. To date, 75 participants have been enrolled and randomized in the trial. We hypothesize that participants in arms 2 and 3 will have increased satisfaction compared to the control group, due to entertainment options provided by the tablet. We also hypothesize that participants in arm 3 will have higher patient engagement due to access to Inpatient myNYP. The primary outcomes of this trial are patient satisfaction (measured using a modification of Telemedicine Satisfaction and Usefulness Questionnaire [TSUQ]), and changes in patient engagement (measured by the Patient Activation Measure [PAM]). The secondary study outcomes include rate of changes to medication orders (e.g., if changes are requested by the patient), changes to length of stay, the rate of refusal by patients to accept medications, and the number of adverse events that occur during the hospital stay. The PAM score is collected at study enrollment, and all other outcomes, including the follow-up PAM score are collected between days 3 and 5 of the hospital admission.

Conclusion: We hypothesize that the use of the inpatient portal will identify and address patients’ information needs, enhance patient engagement, and improve satisfaction.

References
High Level Architecture and Evaluation of Patient Linkages for READY - An Electronic Measurement Tool for Rheumatoid Arthritis

Puneet Mathur, MS, MBA1; Leslie A. Southern, BS1; Aseem Bharat, MBBS, MPH2; Shunchao Wang, MS1; Chris Heckler, BS, PMP1; Omkar Lele, MS, MBA1; Peter J. Embi, MD, MS1; Jeffrey R. Curtis, MD, MS, MPH2
1Department of Biomedical Informatics, The Ohio State University, Columbus, OH
2University of Alabama at Birmingham, Birmingham, AL

Abstract
RhEumatoid Arthritis Disease activity (READY) is an electronic measurement tool for tracking Rheumatoid Arthritis (RA) to help physicians render better clinical care. It collects data using validated instruments to score longitudinal patient reported outcomes (PROs). We here describe the high level architecture of READY tool in a multi institution environment operating on the cloud platform along with evaluation of a mechanism for generating a hashed identifier for patient linkages internally and across external system.

Introduction
Tracking patient reported outcome (PRO) data is increasingly important to clinical research and healthcare improvement. The READY application captures PRO data via various generic and disease-specific instruments. Responses to these instruments allow calculation of scores that measure functional status and RA disease activity. One of the major goals for developing this tool was the ability to store and retrieve patients’ past visit occurrences and visually plot PRO data trends.

Methods and Results
The READY platform functions in a multi institution environment with multiple physicians’ office sites operating under a site ID. The application logic is provisioned by web services deployed over a centralized cloud platform (Figure 1). The challenge of leveraging cloud platform along with fulfilling HIPAA requirements for purposes of patient record linkages was met by creating a one way hashed patient identifier (using a SHA-256 function) from a combined string of selected Personal Identifying Information (PII) data. We evaluated various combinations of PII data elements against approximately 3 million patient records at The Ohio State University Wexner Medical Center in order to achieve the highest positive predictive values and to minimize collisions, or false positives. Results are summarized in Table 1 and show that first 2 letters of first and last name, date of birth, sex, middle initial & physician’s National Provider ID (NPI) number was able to link patients with very few collisions (0.15%).

Conclusion
The architecture and evaluation of patient record linkages has laid the foundation for implementing the application in a multi-site setting. This also opens new opportunities for interacting with external systems easily where linkages between health systems and patient registries (e.g. PCORnet) is desirable using a common hashed patient identifier.

Acknowledgement: This work was supported by NIH P60 AR064172.

References
Clinician Evaluation of Clinical Decision Support Alert and Response Appropriateness

Allison B. McCoy, PhD1,2, Eric J. Thomas, MD, MPH3, Marie Krousel-Wood, MD, MSPH1,2, Dean F. Sittig, PhD3
1Tulane University, New Orleans, LA; 2Ochsner Health System, New Orleans, LA; 3The University of Texas Health Science Center at Houston, Houston, TX

Abstract
Effectively evaluating the appropriateness of clinical decision support (CDS) alerts and responses is critical to improving patient safety through health information technology. Using a REDCap data collection tool, clinicians adjudicated the appropriateness of CDS alerts and responses. Of 397 alerts evaluated, 98.7% were overridden, 38.3% were appropriate, and 97.2% received appropriate responses. Further research is warranted and underway to identify predictors of inappropriate alerts, improve alert specificity, and reduce alert fatigue.

Introduction
Computerized alerts that warn clinicians about drug interactions or provide dosing guidance are commonly implemented forms of clinical decision support (CDS) to improve patient safety. Alert overrides can hinder provider effectiveness and result in adverse patient outcomes. Detailed evaluation of the appropriateness of alerts and clinician responses is necessary; however, few have performed these evaluations.

Methods
Based on a previously described alert evaluation framework, we developed an assessment tool using REDCap to facilitate reviews by clinicians in adjudicating the appropriateness of CDS alerts and responses. We extracted log data from the electronic health record (Allscripts Enterprise EHR) and imported alerts from 100 randomly selected patients who received an alert between January 1 and July 1, 2013 into REDCap to populate these fields prior to clinician reviews. Clinician reviewers then indicated whether the alert and response were appropriate, why the alert was inappropriate (if applicable), and what response(s) would have been appropriate. Each alert was reviewed initially by two clinicians, and a third clinician reviewed alerts when there was a disagreement. We report descriptive statistics on relevant alert, patient, clinician, and medication characteristics.

Results
The clinicians evaluated 397 medication alerts. Of these alerts, 392 (98.7%) were overridden. Clinicians adjudicated 152 alerts (38.3%) as appropriate; 149 appropriate alerts were overridden. Clinicians also adjudicated 386 (97.2%) of alert responses as appropriate; 381 (98.7%) of appropriate responses were overrides, although in 4 cases the medication was not actually ordered, and in other cases clinicians performed additional care, including removing duplicate orders, decreasing doses, ordering monitoring, and documenting awareness of potential interactions. Other causes for appropriate overrides included out of date medication and problem lists, history of patient tolerance despite reported allergies and interacting conditions, and low alert specificity (i.e., alerts about pregnancy contraindications for all women of childbearing age). Trends were similar for all alert types (i.e., drug-drug, drug-allergy, drug-dose, and drug-condition). The median numbers of historical patient alerts, medication orders, and recorded problems were lower for appropriate alerts than inappropriate alerts (16.5 vs 24, 7 vs. 10, and 17 vs. 18 respectively).

Conclusion
We identified low rates of appropriate alerts; most clinician responses to alerts were appropriate, though these were often justifiable overrides or additional provision of care. Poor maintenance of up-to-date patient charts and non-specific alert criteria contributed to inappropriate alerts. High rates of inappropriate alerts that merit overrides by clinicians may have a negative impact on patient care and should be eliminated. Further research is warranted and underway to identify significant predictors of inappropriate alerts, improve alert specificity, and reduce alert fatigue.

Acknowledgments: This work was supported by NLM Grant 1K22LM011430-01A1, a UTHealth Young Clinical and Translational Sciences Investigator Award (KL2 TR 000370-06A1), and NCRR Grant 3UL1RR024148.
Stratification of Risk for Fall Resulting in Hospital Readmission through Medication Side Effects Profiles

Thomas H. McCoy, Jr., MD\textsuperscript{1}, Victor Castro, MS\textsuperscript{1,2}, Roy H. Perlis, MD, MSc\textsuperscript{1}
\textsuperscript{1}Center for Experimental Drugs and Diagnostics, Massachusetts General Hospital, Boston, MA USA; \textsuperscript{2}Partners Research Computing, Boston, MA, USA

Abstract

Fall related injury is a major source of morbidity and mortality. Stratification of fall risk is an important step towards targeted intervention. Epidemiological data suggest that dizziness, syncope, hypotension, and confusion contribute to a patient’s fall risk. Off-target effects of medications with these adverse effects may themselves impart detectable fall risk. Testing this hypothesis in a statewide database, we find that inclusion of medication side effects enhances predictions of hospital readmission due to falls.

Introduction

Falls-related injuries are important sources of morbidity and mortality\textsuperscript{1,2}. Patients’ physiological states, including weakness, dizziness, vertigo, confusion, and hypotension are implicated in the majority of falls\textsuperscript{3}. Off-target effects of medications may induce or exacerbate such states. The desire to target fall prevention efforts has motivated risk stratification attempts; however, to date these attempts have required evaluations that are difficult to translate into practice\textsuperscript{4}. Prior work suggests medication side effects can identify biological consequences of interest\textsuperscript{4,5}. This is appealing, as medication lists are readily available for translation to practice. The work presented here is an attempt to extend this method to enhance predictions of fall risk using side effect profiles in a large, heterogeneous clinical population.

Methods

Diagnostic and sociodemographic data from a statewide claims database were used to derive a logistic regression model of hospital readmission for injury related to fall at 180 days following discharge from the index hospitalization. This model was augmented with a medication side effect feature calculated using previously published medication side effect profiles that were hand annotated as predictive of fall risk\textsuperscript{5,6}.

Results

Of 193,000 identified hospital discharges, 1,900 resulted in readmission following falls. Older age at admission, index admission through the emergency department, higher age-adjusted Charlson Comorbidity Index, and higher calculated side effect profile were independently associated with hazard for fall after adjusting for sex and number of medications prescribed. This model resulted in an AUC of 0.72 for prediction of fall within 180 days of discharge.

Conclusion

Medication side effect profiles available in previously published databases can be used in conjunction with prescriptions and diagnostic claims data to enhance prediction of falls within 180 days of hospital discharge.

References

A Taxonomic Analysis of Programming Errors in Electronic Health Records (EHRs) which Lead to Clinical Decision Support Malfunctions

Dustin S. McEvoy, BS¹, Salman T. Hussain, BSc², Thu-Trang T. Hickman, MPH², Dean F. Sittig, PhD³, Adam Wright, PhD¹,²,⁴
¹Partners Healthcare, Information Systems, Wellesley, MA; ²Brigham & Women’s Hospital, Boston, MA; ³University of Texas Health Science Center, Houston, TX; ⁴Harvard Medical School, Boston, MA, USA

Introduction

With expanded implementation of clinical decision support (CDS) systems in EHRs has come increased concern over patient safety issues that may result when CDS functions improperly.¹ Cataloging and learning from these issues is essential to patient safety and the effectiveness of CDS. While a number of efforts to investigate HIT related safety issues exist, limited work has been done to explore CDS specific issues and their causes.² Given this, we sought to catalog CDS types prone to failure and the programming errors leading to these failures in the EHR.

Methods

We reviewed the Team Coordination System (TCS) database at Partners Healthcare which contains over 100,000 reports cataloging IT related dialog between analysts, developers, engineers and end-users of the Partners EHR. To identify CDS related reports, we queried the database for resolved issues containing terms such as “alert” or “DDI”. Reports related to CDS were analyzed to determine the CDS type and programming error involved. CDS type was classified using previously developed taxa³ while programming error type was classified using a subset of the taxa presented by Ko and Myers.⁴ Reports were classified with a single programming error type and up to 2 CDS types.

Results

Of the 4,941 reports meeting our search criteria, 1,500 have been reviewed to date and 93 (6%) distinct reports have been deemed related to CDS. The majority of these reports have involved point of care alerts and reminders (63%), while reports related to medication dosing support (18%), order facilitators (12%) and workflow support (12%) were also frequently observed. The types of programming errors observed varied more considerably. Among the reports where programming error type could be determined from the dialogue, omission errors (24%), trivial typos (19%) and imperfect design specifications (18%) were most common. Programming error type varied with CDS type, but there was insufficient data to determine a clear relationship between the two.

Conclusion

This is the first catalog of the types of CDS prone to failure and the programming errors leading to these failures in the EHR. Initial observations indicate that programming errors are major barriers to the proper implementation of CDS. Various programming error types were observed across the reports, suggesting that careful CDS specification development and intervention testing are essential in preventing patient safety issues. While the majority of the reviewed reports were related to point of care alerts and reminders, this may be due to bias introduced by the search terms used to extract these reports, which focused primarily on descriptions of this type of CDS. We are expanding on this work to further investigate causes, types, and inter-relationships that result in CDS malfunctions.

References

Partnering to Develop a Service-based CDS System for Public Health Reporting Specifications

Sunanda McGarvey, CBAP, CSM; Denisha N. Abrams, Janet Hui, MPH; Laura Conn, MPH; Catherine Staes, MPH, PhD; and the RCKMS Team

1Northrop Grumman Information Technology, Atlanta, GA., 2Council of State and Territorial Epidemiologists, 3Centers for Disease Control and Prevention, 4Department of Biomedical Informatics, Univ. of Utah, Salt Lake City, Utah

Background
In the US, selected conditions must be reported to public health authorities by physicians, hospitals, and laboratories to control disease and monitor population health. Currently, reporters must manually review various sources to determine reporting criteria, and often clinicians are unaware of reporting requirements (1). The Reportable Condition Knowledge Management System (RCKMS) project, supported by the Centers for Disease Control and Prevention and the Council of State and Territorial Epidemiologists, is focused on developing an infrastructure to allow public health authorities to author their logic and present reporting criteria in a human-readable and electronic format. The objectives of the project is to: a) describe variation in reporting specifications for 4 reportable conditions across 11 jurisdictions; b) define logic to represent the reporting specifications as machine-processable rules in a public health decision support system; and c) collaborate with a healthcare enterprise to use the public health decision support system to determine if a case is reportable to public health.

Methods
We convened a workgroup of public health experts from seven states (CO, DE, IL, NY, UT, VA & WA) and four counties (Houston, New York City, Southern Nevada & San Diego), subject matter experts, and experts from Intermountain Healthcare, and HLN Consulting, LLC weekly from Nov 2014 through July 2015. The workgroup assessed variability in reporting criteria, identified areas for harmonization and simplification, and proposed a default set of logic for each condition. In a parallel effort, vocabulary subject matter experts convened twice a week to draft standardized value sets and rules algorithms to support the default logic. To manage the reporting specifications and generate machine-processable rules based on the logic, the team worked with HLN Consulting, LLC (the Clinical Decision Support [CDS] implementers). Informaticians from Intermountain Healthcare engaged as an entity required to report to public health. They provided insight from the inpatient, outpatient, and laboratory perspectives. Weekly meetings were held with HLN Consulting, LLC and Intermountain to define implementable technical approaches that supported public health reporting requirements and moved existing manual text-based processes to automated machine-based processes. An interface for public health professionals to author reporting logic is being designed, and the logic implemented using the service-based CDS infrastructure will be tested and validated.

Results
Among the jurisdictions involved, we identified criteria with varying levels of specificity and different terms to represent common criteria. For tuberculosis alone, thirteen variations of clinical criteria could be rolled up under 6 common criteria. Similarly, for pertussis, the public health practitioners were able to harmonize and agree on a proposed set of logic that includes sixteen criterion (i.e., clinical, lab, epi, demographic criteria) grouped into 6 sets of logic. Similar results were achieved from harmonization of Chlamydia and Lead criteria. HLN, LLC is translating the sets of logic into computable rules that Intermountain will access through a service-based CDS infrastructure to determine if a report should be sent to public health, to which jurisdictions it should be sent, and what it should data elements it should include. The poster will include up to date progress on the development and testing of RCKMS.

Conclusion
Variation does exist among jurisdictions, however jurisdictions can reduce variability and logic can be implemented in a service-based CDS system. Collaboration across public health, and with healthcare, is vital to making automated case reporting a reality.
Visualization of Laboratory, Vital, Precaution and Patient Status Data to Optimize Time on Task and Use Related Hazards

Ray McKelvy,1 Nathan C. Gonzalez,1 Karol L. Fleming JD, MBA,1 William B. Webb BSPH,1 MBA, Robert L Shapiro MD, MSCI1,2

1 MEDarchon, Inc. 113 29th Avenue South, Nashville, TN 37212
2 Cardiovascular Division, Washington University in St. Louis School of Medicine, 660 S. Euclid Avenue, St. Louis, MO 63110

Background: Handoff incorporates a vast amount of data which is subject to use related hazards.1 Negating use related hazards and reducing time on task are significant factors for adoption. This study sought to test the time to comprehension, use related hazards and user satisfaction associated with various methods of visualizing handoff data.

Methods: Doctors and nurses with at least two years of clinical experience were selectively recruited based upon their role, specialty and patient population. After informed consent, subjects were presented with data visualizations for a sick patient and a healthy patient and asked to identify the healthy patient as quickly and accurately as possible. Visualizations tested included four chronological displays of lab and vital data, four displays of current and trend data ways of organizing laboratory findings, and two ways of presenting precautions and patient status (condition, airway, ambulatory, etc.). The order of the visualizations and placement of the sick vs. healthy patient was randomized. For each category, an electronic medical record visualization served as the control. A study administrator recorded and graded responses. Descriptive statistics and the Z test were used to summarize the findings. Results: 23 of 24 recruited subjects completed the study. Table 1 shows the mean time to comprehension, total errors and user preference for various graph formats tested, while Table 2 shows the same for the “graph-plus” format. There was no statistical difference in time to comprehension for organizing labs by the test type (p=0.651) or result type (p=0.412). Visualizing patient precautions (3.90 vs. 4.88 s, p=0.206) and status (4.01 vs. 7.39, p=0.053) with icons trended toward decreasing time to comprehension but did not reach significance. 78% and 61% of participants preferred the icon visualizations for precautions and status, respectively. Conclusion: This study has shown that the method of visualizing handoff-related data can reduce time to comprehension without negatively impacting use related hazards.

Table 1: Time to Comprehension, Errors and User Preference for Graph Format

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean (s)</td>
<td>mean (s)</td>
<td>p</td>
<td>mean (s)</td>
</tr>
<tr>
<td>Single Factorial Time</td>
<td>3.03</td>
<td>2.35</td>
<td>0.091</td>
<td>2.14</td>
</tr>
<tr>
<td>Multi-Factorial Time</td>
<td>5.62</td>
<td>6.13</td>
<td>0.555</td>
<td>4.73</td>
</tr>
<tr>
<td>Errors</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Preference</td>
<td>30.4%</td>
<td>26.1%</td>
<td>26.1%</td>
<td>17.4%</td>
</tr>
</tbody>
</table>

Table 2: Time to Comprehension, Errors and User Preference for “Graph-Plus” Format

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean (s)</td>
<td>mean (s)</td>
<td>p</td>
<td>mean (s)</td>
</tr>
<tr>
<td>Single Factorial Time</td>
<td>2.88</td>
<td>2.66</td>
<td>0.575</td>
<td>2.25</td>
</tr>
<tr>
<td>Multi-Factorial Time</td>
<td>3.12</td>
<td>3.09</td>
<td>0.928</td>
<td>2.60</td>
</tr>
<tr>
<td>Errors</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Preference</td>
<td>8.7%</td>
<td>4.3%</td>
<td>17.4%</td>
<td>69.6%</td>
</tr>
</tbody>
</table>

Born to Lose (the Call): Date of Birth Errors in Patient Identification in an Automated Adverse Drug Reaction Call System

Jeffrey Medoff, BA1; Alejandra Salazar, PharmD1; Elissa V. Klinger, MS1; Japneet Kwatra, BS1; Mary Amato, PharmD, MPH1,3; Patricia Dykes, PhD1; Jennifer S. Haas, MD1,2; David W. Bates, MD, MSc1,2; Gordon Schiff, MD1,2

1 Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA; 2 Harvard Medical School, Boston MA; 3 MCPHS University, Boston MA

Introduction

Interactive Voice Response Systems (IVRS) have emerged as a method of reaching out to large numbers of patients in a cost efficient manner.1 We implemented IVRS for proactive identification of symptoms that may represent adverse drug reactions (ADRs). IVRS usability may be complicated by technological failures, operational challenges, patient acceptance or confusion with using this technology. One challenge, ensuring confidentiality, required that IVRS correctly identify the targeted patient. Using date-of-birth (DOB) verification is one approach we have used, but such a method can introduce unexpected complexities. After encountering various issues using DOB patient identification, we analyzed our experience to better understand reasons for IVRS DOB failures with the aim to improve system user-friendliness and response rates.

Methods

We implemented a randomized controlled trial to screen patients for adherence and ADRs associated with newly prescribed medications. After agreeing to participate, patients were asked to verify their DOB. Instances of unsuccessful verification (“DOB errors”) triggered an alert to research staff members, who manually reviewed the recording of each encounter. Errors were coded as one of three types: 1) patient response did not match DOB in the electronic health record; 2) response was inconsistent with format required; or, 3) response not correctly captured by voice recognition (background noise, delayed response). Some errors represented multiple types. DOB errors resulted in call termination (to protect against surveying unverified patients). Initially, DOB verification requested the entire date, but we later [pause]. Because high DOB-error rates continued, we reverted back to whole date format, allowing calls to proceed regardless of IVRS-DOB verification and instead had research staff manually confirm DOB for each patient.

Results

Overall DOB error rate was 34.0% (235 total errors for 691 patients who agreed to participate) with the majority (61.7%) due to inconsistent format. Surprisingly, patients experienced DOB errors at similar rates when prompted for whole date or sequentially (19.9% and 22.9%, respectively). Finally, when calls were allowed to continue without IVRS DOB verification, DOB error rates increased (44.1%). DOB error rates for Spanish-speaking patients (57.6%) were higher than English-speaking patients (30.1%).

Conclusion

We found high rates of technical DOB errors disqualifying large numbers of willing participants. The leading cause of DOB-errors was patient difficulty in comprehending and adhering to structured IVRS DOB verification format, with non-English speaking patients experiencing higher error rates. Restructuring IVRS prompts to parse birthdates into separate month, day, and year fields unexpectedly did not improve data capture. Requiring multiple attempts did result in increased success rates but not sufficiently to overcome this problem. More thorough pre-implementation piloting is warranted to ensure effective and equitable usability of such systems.

References


This study was funded by grant # 5U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ).
Exploring Gaps of Family History Documentation in the EHR

Saeed Mehrabi, PhD Candidate¹, Maya S. Safarova, MD, PhD², Iftikhar J. Kullo, MD², Hongfang Liu, PhD¹
¹Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; ²Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA

Abstract

In the era of precision medicine, accurately identifying familial conditions is crucial for providing target treatment. In this work, we studied the documentation of family history of stroke in clinical notes and compared it with patient-provided questionnaire (PPI). The agreement between patient-provided information and clinical notes on absence of family history information in PPI was higher compared to its presence.

Introduction

It has been shown that a wide range of adult conditions such as diabetes, cardiovascular diseases, Alzheimer’s and cancers have hereditary roots. Accurate family history information can be very helpful in precision medicine that tailors the treatment to the individual characteristics of patients.

Material and Methods

We have selected 27998 with family history of stroke or transient ischemic attack (TIA) and also retrieved their relevant PPI forms. We used our previously developed rule based NLP algorithm to extract family history information and associate family members and diagnoses¹.

Results:

The agreement between PPI and medical records was evaluated by randomly selecting 25 patients who reported having family history of stroke/TIA in their PPI form and 25 who did not. The positive and negative agreements are 0.653 and 0.7384 respectively.

Table 4. Positive and negative agreement measurements of stroke mentions in PPI forms and medical records

<table>
<thead>
<tr>
<th>Medical Records</th>
<th>Patient Provided Information</th>
<th>Presence</th>
<th>Absence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence</td>
<td>16</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Absence</td>
<td>12</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>

We also evaluated our NLP performance by randomly selecting 100 sentences. There were 19 incorrect instances in 100 sentences resulting in accuracy of 81% of the NLP system. Majority of these errors (15 out of 19 or 79%) were due to incorrect family member associations.

Conclusion

We have identified that there are significant gaps in the family history documentation for making the correct diagnosis of familial conditions. Adequate family history information including detailed information such as affected family members and their social history, age of disease onset, and specific information regarding the disease in question are crucial.

References

Development of a Novel Application for Home Management of Chronic Low Back Pain

Brittany L. Melton, PhD, PharmD¹, Hannah J. Kevern², Jenna M. Stones², Neena Sharma, PT, PhD²

¹University of Kansas School of Pharmacy, Lawrence, KS; ²University of Kansas Medical Center, Kansas City, KS

Abstract

The objective of this project is to develop and test an iPad application for home-based therapy of chronic low back pain. Up to twenty patients will utilize the application instead of in-person physical therapy for six weeks. The application educates patients on pain physiology, fear avoidance, exercises, and pain management options. Patients will complete the Computer System Usability Questionnaire to assess usability, as well as questionnaires to evaluate changes in fear avoidance and pain control.

Introduction

Low back pain (LBP) is the second leading cause of disability in the United States, costing an estimated $560-635 billion annually, and the incidence is increasing.¹ ² Exercise therapy is the standard of care for chronic LBP but less than half of patients have been prescribed exercise, and only 30% have seen a physical therapist.³ ⁴ At the same time, exercise therapy does not address the psychological factors associated with chronic pain which can negatively impact pain management. Combining exercise therapy with pain education in a home-based program could improve patient outcomes and improve access for rural patients. An iPad may be useful in facilitating physical therapy access regardless of patient location and providing comprehensive care to manage LBP. The objective of this project is to develop and test an iPad application for home-based therapy of chronic low back pain.

Methodology

The patient education and exercise program is an iPad-based application. The initial prototype was developed by clinician study investigators by incorporating pain education and strategies for fear avoidance, stretching, strengthening and aerobic exercises, and relaxation techniques into a one hour session. The application reflects the frequency and duration of traditional physical therapy sessions, by providing three sessions a week for six weeks. After the first two weeks participants will have a virtual meeting with a study investigator to ensure their proper understanding of the exercise program and accessibility to the application content. Up to twenty adult participants with chronic LBP, will be recruited for the study. Both qualitative and quantitative measures will be collected. Participants will complete questionnaires assessing their depression, fear avoidance, disability, and pain at baseline and every two weeks during the study. Additionally participants will complete a telemedicine satisfaction questionnaire and the validated Computer System Usability Questionnaire after completing the program. Both usability results and participant reported changes in pain and fear will be evaluated to guide further application development. Participant comments during the virtual meeting will also be qualitatively analyzed for emergent themes to further guide development. This study was approved by the Institutional Review Board.

Results

Data collection is currently in process.

Conclusion

Testing is ongoing to design and optimize an iPad application to facilitate exercise and education for chronic LBP patients. Additional conclusions will be made once preliminary data has been collected and analyzed.

SCILHS Data Mart Creation Plugin

Michael E. Mendis, Janice Donahoe, Lori C. Phillips MS, Vijay A. Raghavan MS, Alexander Turchin, MD MS, Jeff Klann PhD, Shawn N. Murphy MD PhD

Partners Healthcare Inc., Massachusetts General Hospital, Harvard Medical School, Boston

Abstract

The Scalable Collaborative Infrastructure for a Learning Health System (SCILHS, pronounced “skills”) is a network of 11 health centers across the United States that will cover over 8 million patients. SCILHS is a Clinical Data Research Network (CDRN) in the Patient-Centered Outcomes Research Institute’s PCORnet.

Introduction

SCILHS has developed an Informatics for Integrating Biology and the Bedside (i2b2) information model that represents the PCORnet Common Data Model (CDM). This information model consists of an i2b2 ontology/terminology and process for mapping local data elements to the CDM. The ontology mapping is done without changing the underlying imported data. This approach highlights i2b2’s ability to separate the CDM from both information model and the underlying data format.

The sites will be able to programmatically generate i2b2 data marts that are based on the PCORnet CDM format, which will help assist in the patient clinical research. The DSD (Disease Specific Data mart) creation script assists the sites with creating a new i2b2 data mart based on an existing i2b2 query patient set.

For instance if a PI wants to do research on weight management, the first step is to run a query in i2b2 to find their set of patients. Using the plugin the PI can create a new disease specific data mart with only the patients in their defined i2b2 patient set. All of the data associated to those patients will be copied into the new data mart along with the ontology. The plugin has the flexibility to create either an Oracle or SQL Server script that contains the appropriate SQL statements that are needed for a DB to create the new data mart. Once the data mart has been created the PI can use the CT viewer within the i2b2 Workbench to manually review each patient to determine if they are suitable for the study.

The CT viewer provides the PI a friendly, user interface for rapid line-by-line viewing of patient attributes, and the ability to drill deep into an individualized view that can appear very similar to what an electronic medical record could produce.

Conclusion

The SCILHS data mart plugin assists the PI in fine-tuning a patient cohort for their study, by creating a data mart that contains only those patients who meet the specified criteria. Once they have this set, they can use the Clinical Trial (CT) viewer to determine if the patient is suitable for the study.

This work was (partially) supported through a Patient-Centered Outcomes Research Institute (PCORI) Award CDRN-1306-04608 (282364.5077585.0007)
Standardization of Ask At Order Entry Questions: A Prudent Question is One-Half of Wisdom

Riki Merrick, MPH\(^1\), Cynthia Johns, MLS\(^2\), Pamela Banning, BSMT\(^3\)
\(^1\)Vernetzt, LLC, Sausalito, CA, \(^2\)LabCorp, Burlington, NC, \(^3\)3M Health Information Systems, Murray, UT

**Problem:** The United States federal mandates for achieving Meaningful Use (MU) cause numerous work groups in the industry to discuss how to best represent actual content with standardized terminology. During the development of the Laboratory Orders Interface (LOI) and the electronic Directory of Services (eDOS) implementation guides in the Standards and Interoperability Framework’s (S&I Framework) laboratory related initiatives, the need for guidance, expansion and harmonization of Ask at Order Entry (AOE) questions became evident. Examples of such questions include: collection duration and volume of 24 hour urine for chemistries, and pregnancy/gestation questions for maternal serum integrated screening assays.

**Objective:** The primary goal is to provide a suite of harmonized standards for electronic messaging of laboratory data for the US realm for inclusion in MU regulations. A secondary goal is the reduction of variations in how the same AOE question is asked, for example, History of Down Syndrome (LOINC 53826-4) versus Prior Pregnancy with Down Syndrome. Use cases considered for this project were ambulatory laboratory test ordering as well as Public Health reporting, such as pediatric lead level reporting. Engagement of subject matter experts who help with consolidating duplicate AOE questions into a standard format for representation as a single concept provides further harmonization and enhanced interoperability. In addition, this project will establish a review process prior to submission for standard codes when new AOE questions are needed.

**Methods:** During the course of LOI and eDOS implementation guide development, commonly used AOE questions were collected from several national laboratories, Public Health laboratories and Public Health agencies. AOE questions were consolidated and standard codes were assigned from the Logical Observation Identifiers Names and Codes (LOINC\(^{®}\)) database, maintained by the Regenstrief Institute, where appropriate codes existed. AOE questions where the required information could be communicated in other parts of the HL7 message were also identified. Several of the AOE questions were asking the same questions in different ways and this duplication creates confusion and hinders semantic interoperability. The entire AOE collection was shared with the Laboratory Messaging Community of Practice (LMCoP), a forum comprised of laboratory and standards experts from state and federal Public Health laboratories, national clinical laboratories, the National Library of Medicine and professional organizations, whose purpose is to resolve lab related issues from a laboratorian’s viewpoint. The LMCoP, acting as a conduit to lab related professional organizations, provided its recommendations for subsequent review by appropriate laboratory domain content experts from the American Society for Clinical Pathology (ASCP), the Association for Molecular Pathology (AMP), and the College of American Pathologists (CAP) for completeness and proper LOINC\(^{®}\) code assignment, as well as identification of a preferred single concept representation where overlap existed.

**Results:** The current iteration of AOE questions contains 131 questions, a 37% reduction from the 210 originally collected. Fifty eight (58) new terms have been submitted to the Regenstrief Institute; in addition twenty three (23) existing LOINC\(^{®}\) terms were revised, removing trial status or survey-specific method information as result of this review. Specific test-related AOE questions will be grouped into LOINC\(^{®}\) panels in the coming releases.

**Conclusion:** The curated list of AOE questions, properly mapped to LOINC\(^{®}\) terminology, has been published in the eDOS Implementation Guide and is available to laboratories when implementing electronic data exchange. This project also established the review process for future new AOE questions for proper representation and evaluation of possible duplication, prior to submission for LOINC\(^{®}\) assignment by the Regenstrief Institute.

**References**

Intelligent Home Risk Monitoring Solution Enable Post Acute Care Surveillance

Hoda Moghimi, PhD, MSc., BEng1, Nilmini Wickramasinghe, PhD, MBA, GradDip, MgtSt, BSc.1,2, Jonathan L. Schaffer, M.D., MBA.3
1Epworth HealthCare Melbourne, Victoria, Australia; 2Epworth HealthCare and Deakin University, Melbourne, Victoria, Australia; 3Cleveland Clinic, Ohio, USA

Introduction

The advancing age of baby boomers, coupled with increased life spans, has led to a significant increase in the number of senior citizens in many countries. These populations are projected to significantly impact current and future healthcare resources. Providing care for this population in the acute care setting is only one aspect of the total care package that needs to be addressed. For those having been in the acute care setting for either medical treatment or following procedural based therapies, the discharge to home often provides an opportunity to continue the post acute care monitoring to ensure that complications or readmissions do not occur. Monitoring care and providing guidance and medical management at home will offer patients, families, facilities and providers with the opportunity to ensure recovery and return to a healthy steady state. To explore this issue further, the following examines the possibilities for monitoring post-operative clinical functions in the context of Total Knee and/or Total Hip Arthroplasty (TKA/THA). Specifically, this research in progress serves to proffer a conceptual model that can then guide a randomised clinical trial to test the presented Intelligent Home Risk Monitoring (IHRM) Solution.

Proposed Intelligent Home Risk Based Monitoring Solution

Decision-making regarding the treatment and surgery for patients under TKA/THA is especially multi-faceted and complex, including for example the clinical condition of the patients, their age and size. The decision to athroplasty treatments with either drugs, or surgery, or a combination of both depends on a large number of factors1. The decision making process in TKA/THA can be divided into six broad phases, as awareness program, outpatients clinic, pre-operative evaluation, day of surgery, post operative inpatient and post operative rehabilitation2. In the current process, nearly all patients receive post operative physical therapy and rehabilitative services in the hospital and despite the cost effectiveness of TKA/THA, rehabilitation costs in hospitals associated with these surgeries also place significant burdens on healthcare systems. Hence, it is understood that post operative home risk monitoring can be proposed to all patients who are identified in low risks or some patients with moderate post operative risks, through adopting a home based monitoring technology, as demonstrated in Figure1. The solution major functionalities include providing cause, effect diagrams and influence diagrams, showing the casual influences or relationships among variables and outcomes. Risk probability and impact assessment and risk categorization are further main capabilities of the solution.

Therefore, benefits of using the IHRM solution at home by both senior citizens and professional care providers can be summarised, as primarily clinical for optimal quality of care and access since these patients can now enjoy a post operative home care in a comfortable environment, early detection of post operative risk factors and also enhancing their autonomy through self-monitoring post operative conditions but also as economic benefits through saving rehabilitation costs in hospitals, logistics costs due to regular clinical visits and even environmentally friendly.

Conclusion

This research in progress has far reaching implications especially in light of the challenges facing many healthcare systems of an ageing population coupled with escalating healthcare costs and increases in chronic diseases as well as workforce shortage issues.

References

Improving Evidence-Based Migraine Management in VA Primary Care Clinics by Utilizing Informatics Tools

Amir Mohammad, MD, MPH, FACPM, Hamada Hamid, DO, Cynthia Brandt, MD, MPH
VA Connecticut Healthcare System, Yale University School of Medicine

Abstract
Migraine headaches are a common cause of impairment and lost employment productivity; people with migraines miss on average four days of work a year and cost the U.S. an estimated $13 billion a year. Veterans are particularly vulnerable to developing migraines because of their relatively high rate of traumatic brain injury. Having an immediate access to the evidence-based guidelines as a tool will assist in educating our clinicians to effectively manage migraine.

Introduction
Veterans are vulnerable to developing migraine headache. A study found that migraine headaches are more prevalent in service members who were deployed to Iraq. Another study found that the prophylactic medications to treat migraines were greatly underutilized in primary care settings. This study also highlighted the importance of enhancing the knowledge of general practitioners regarding the diagnosis of migraine, and to increase awareness of the headache guideline.

Methods & Results
Our baseline survey noted that the primary care clinicians were not consistently utilizing the evidence-based guidelines as only 65% reported utilizing American Academy of Neurology (AAN) guidelines when managing patients with migraine headaches. This survey also indicated that 75% of clinicians were prescribing anti-migraine prophylaxis and almost all of them (91%) were referring these patients to neurology clinics. This quality improvement project identified several non-value added steps. Since this project is currently being conducted in one of the primary care clinics, final results will be presented later.

Conclusion
Providing patient-centered care to veterans with migraine headaches require removal of non-value added steps in clinics and implementation of electronic tools. Utilizing the orders module in Computerized Patient Records System (CPRS), we will develop a Migraine Reminder Template and hyperlinks to Clinical Guidelines for Migraine.

Having an immediate access to the evidence-based guidelines as a tool will greatly assist in educating our clinicians, trainees, and nursing staff; and will subsequently promote timely and effective migraine management in VA Primary Care Clinics.

References
Assessing and Simulating Scheduling Processes in Community Health Centers

Iman Mohammadi, MS¹, Ayten Turkcan, PhD², Tammy Toscos, PhD¹,³, Amy Miller, MA¹, Kislaya Kunjan, MS, MBA¹, Brad N. Doebbeling, MD, MSc¹
¹Department of BioHealth Informatics, Indiana University, Indianapolis, IN; ²Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA; ³Parkview Research Center, Fort Wayne, IN

Problem. Community Health Centers (CHCs) are safety-net clinics providing primary care for underserved populations. Improving access to care is critical for establishing health equity, increasing quality, and improving health outcomes. In our ongoing research to improve access in CHCs, we found patients are concerned about long waiting times to get an appointment, not getting appointments at convenient times, and not getting appointments with preferred providers. Clinics face appointment scheduling challenges including provider shortage, limited provider availability, multiple patient visit types, appointment no-shows and cancellations.

Purpose. Effective scheduling processes can reduce clinic no-show rates and patient waiting time while improving continuity of care and overall clinic performance. We sought to develop a computer simulation model to assess and simulate the scheduling processes in CHCs, and provide a decision making tool for clinic managers to analyze the impact of a modified open access scheduling system, where some provider capacity is allocated for same-day appointments.

Methods. In this poster, we will describe the methodology used to develop the simulation model using clinical and operational data collected from CHCs. The data requirements, methods used for data collection, data analyses needed to represent the clinical practice realistically, and methods used for validation of simulation results will be characterized. Data requirements for the simulation model include provider characteristics, patient characteristics, appointment types, visit frequencies, and scheduling methods. Structured questionnaires and interviews were used to gather data about type of services offered in each clinic, number of providers, nurses and staff, access modalities, process to make an appointment, scheduling method and horizon, and enrollment visit processes. Clinic managers, staff, quality assurance directors, schedulers, financial advisors, nurse managers, call center staff and front desk staff were the key respondents. Workflow observations are necessary to fully understand the scheduling process in each CHC. The clinic staff working at the front desk (check in/check out), call center, scheduling and enrollment stations were observed to map the scheduling processes. Data from questionnaires and observations will be used to create a representation of patient flow and develop the scheduling algorithm. EMR data will be used to build patient population characteristics, and visit frequencies. No-show probabilities and cancellation rates will be estimated using statistical methods such as logistic regression. The simulation model will be validated by comparing the simulation output results with the CHC performance measures. The performance measures may include waiting times for appointments, number of patients seen by their primary care providers, and provider utilizations.

Preliminary Results. We have observed the scheduling processes in seven community health centers across state of Indiana. Our observations show that the scheduling process changes according to different patient types (new vs. established). Acute, non-acute, follow up and insurance enrollment (e.g. Medicaid and Medicare) visits are typically the four main appointment types. An agent-based simulation model was developed in AnyLogic in our earlier study¹. Here, the scheduling algorithm is updated based on our observations (e.g. scheduling enrollment visits for both established and new patients; scheduling enrollment and provider appointments on the same day; scheduling of patients with complicated conditions with physicians as oppose to nurse practitioners; and providing access through walk-in hours). For validation of the simulation model, the performance measures important for the CHCs are included in the model.

Conclusion. This project is part of a 3-year study funded by PCORI to understand and improve access to healthcare in Indiana through collaboration with seven community health centers in the state. Questionnaires and interviews for understanding overall operations of the partner clinics, workflow observations and EMR data analysis can be used to build the simulation model to identify effective scheduling processes and test alternate strategies to improve timely access to care.

References

1598
Usability of a phenotype builder prototype and lessons learned for the design of phenotyping tools

Enid Montague, PhD1; Jie Xu, MS1; Luke V. Rasmussen1; Joshua C. Denny, MD, MS2; Guoqian Jiang, MD, PhD3; Richard C. Kiefer3; Huan Mo, MD, MS2; Jennifer A. Pacheco1; Peter Speltz2; William K. Thompson, PhD4; Jyotishman Pathak, PhD1

1Northwestern University, Chicago, IL; 2Vanderbilt University, Nashville, TN; 3Mayo Clinic, Rochester, MN; 4NorthShore University HealthSystem, Evanston, IL;

Abstract

The use of electronic health records (EHRs) for the development of phenotype algorithms offers much potential, but is hindered in part by the lack of portable, standardized phenotype definitions. One solution is the creation of novel informatics tools to facilitate algorithm development. Here we develop design guidelines from potential end users of phenotyping software using qualitative methods. Results are presented as themes to inform the design of future software systems.

Introduction

Phenotyping is the process of systematic collection and analysis of phenotypic data.1 The implementation of electronic health records (EHRs) on the national level has a potential to tremendously scale the phenotyping process for clinical research through the creation of algorithms and tools to extract meaningful information from EHRs.2 However, one challenge is the lack of validated tools to author and apply standardized research phenotype algorithms across different sites and EHR systems accurately and efficiently. Towards this goal, participatory design enables designers and developers to learn user needs and preferences early in the design process through a workshop of diverse stakeholders. The results are early design ideas that can lead to the development of more accurate user requirements for research phenotype algorithm creation.

Methods

Following the process of participatory design, we invited 15 potential end users at Northwestern University and Mayo Clinic to use a prototype tool called “Phenotype Builder” prior to the design session. All participants provided informed consent to participate. Each participant was given access to the tool, two use cases and asked to create two phenotype algorithms. The use of the tool was recorded using screen capture software. After they completed the cases to the best of their abilities, all participants participated in a focus group to discuss their needs regarding the design of a future phenotyping tool.

Results

Most participants had experience with existing tools such as the Measure Authoring Tool (for creating electronic Clinical Quality Measures) and i2b2. Some had very little experience with tools, but had interest and/or experience in algorithm development. The primary usability requirements that emerged from the group were to (1) Present the algorithm in both graphical and technical form. (2) Provide a feedback system for the outcome of the algorithm. (3) Allow for flexible definitions. Since the main stakeholders of this platform are researchers, the definitions should be able to be extremely flexible for exploratory research design. (4) Provide built-in definitions. Electronic clinical quality measures usually have precise definitions of patient cohorts, however, this is less likely to be true in research. It could be useful to integrate the precise definitions from the quality-based work so that the researchers can cite directly if they want. (5) Develop detailed documentation for the platform. (6) Include search functions so that codes in standard vocabularies (ICD, RxNORM, LOINC, etc.) are searchable, and provide guidance toward which concepts are actually used.

Discussion

Participants provided requirements for the design of phenotyping tools that were related to the algorithm building activities and general usability. Developing a more useable system for building phenotyping algorithms can contribute to more robust research and clinic practice.

Acknowledgement. This work has been supported in part by funding from the NIH (R01-GM105688).

References

Using Natural Language Processing to Facilitate Medical Record Abstraction in Epidemiological Studies

Carlton Moore MD MS1,3,4, Kyle Shaffer BA3, Anna Kucharska-Newton PhD MPH2, Stephanie Haas PhD,3,4 Gerardo Heiss MD PhD2

1University of North Carolina School of Medicine; University of North Carolina at Chapel Hill; 2School of Public Health; 3School of Information and Library Sciences; 4Carolina Health Informatics Program

Introduction. The Atherosclerosis Risk in Communities (ARIC) study1 conducts ongoing surveillance of hospitalized cardiovascular health events and death in 4 communities in the United States (NC, MI, MN and MD). Diagnostic criteria for heart failure (HF) has been manually abstracted from medical records since 2005,2 including the presence of symptoms consistent with HF decompensation (new onset or worsening shortness of breath, edema, paroxysmal nocturnal dyspnea, and orthopnea) during patients’ hospitalizations. The manual chart abstraction process has high repeatability under a stringent quality control protocol, but is time consuming and costly. The goal of this study is to develop and test natural language processing (NLP) tools to extract information on complex symptoms of HF from free-text electronic medical records.

Methods. We used c-TAKES (clinical Text Analysis and Knowledge Extraction Tool), an open-source NLP software package3, to identify the presence of symptoms of new onset or worsening shortness of breath, edema, paroxysmal nocturnal dyspnea, and orthopnea in de-identified free-text electronic medical records from the ARIC study. We then created a program, written in Python, that processes the c-TAKES output to negate relevant concepts not consistent with new onset or worsening symptoms. For example, the shortness of breath concept in the text “the patient has chronic shortness of breath” is negated because the symptom is not new onset or worsening. Using an initial corpus of 112 medical records (representing 112 unique hospitalizations) adjudicated by ARIC nurse chart abstractors as the “gold” standard, we evaluated the recall and precision of the NLP system. We also investigated instances of discordance between the “gold” standard assessment and NLP results in order to identify areas for NLP algorithm improvements.

Results. The recall and precision of NLP vs. ARIC abstraction in identifying HF symptoms in the free-text corpus of 112 medical records is shown in Table 1. Recall was 100% for shortness of breath, paroxysmal nocturnal dyspnea and orthopnea; and 98% for edema. A review of instances in which NLP identified HF symptoms that ARIC abstractors had not (i.e., false-positives) suggests that NLP may be more sensitive than ARIC nurse chart abstractors (Table 2). For example, in one instance the medical record stated the patient “Has had runs of V-tach (ventricular tachycardia) overnight with shortness of breath”. This was correctly identified by NLP as new onset or worsening shortness of breath, but not by ARIC nurse abstractors. The precisions (Table 1) shown in parentheses are corrected for instances in which symptoms identified by NLP were overlooked by abstractors.

Table 1. NLP Performance Characteristics (n = 112)

<table>
<thead>
<tr>
<th>ARIC HF Variable</th>
<th>Recall</th>
<th>Precision</th>
<th># of patients in which symptoms identified by NLP were overlooked by abstractors</th>
</tr>
</thead>
<tbody>
<tr>
<td>New onset or worsening shortness of breath</td>
<td>100%</td>
<td>76% (91%)</td>
<td>13</td>
</tr>
<tr>
<td>New onset or worsening edema</td>
<td>98%</td>
<td>53% (66%)</td>
<td>11</td>
</tr>
<tr>
<td>Paroxysmal nocturnal dyspnea</td>
<td>100%</td>
<td>64% (73%)</td>
<td>1</td>
</tr>
<tr>
<td>Orthopnea</td>
<td>100%</td>
<td>81% (90%)</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2. Examples in which symptoms identified by NLP were overlooked by ARIC abstractors

<table>
<thead>
<tr>
<th>ARIC HF Variable</th>
<th>Relevant Free-Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>New onset or worsening shortness of breath</td>
<td>Has had runs of V-tach overnight with shortness of breath. EKG overnight was unchanged. Troponins continue to be trended.</td>
</tr>
<tr>
<td>New onset or worsening edema</td>
<td>Her sx are c/w volume overload, severe valvular disease. Bilateral lower extremity edema 2+</td>
</tr>
<tr>
<td>Paroxysmal nocturnal dyspnea</td>
<td>Frequent PND and pedal edema</td>
</tr>
<tr>
<td>Orthopnea</td>
<td>Does endorse some orthopnea</td>
</tr>
</tbody>
</table>

Discussion. We developed an NLP application that has close to 100% recall for identifying HF symptoms in free-text electronic documents. The NLP algorithm is being refined to further improve its performance characteristics and to extract additional ARIC HF symptoms embedded in free-text medical records. Our preliminary results indicate that, in some instances, NLP may be more sensitive than manual abstractors; and our results suggest that use of validated NLP technology holds the potential for improving the cost-effectiveness of epidemiologic surveillance.

1 “The Atherosclerosis Risk in Communities (ARIC) Study.”
2 Rosamond et al., “Classification of Heart Failure in the Atherosclerosis Risk in Communities (ARIC) Study.”
3 Savova et al., “Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES).”
The Everyday Practice of Health for Mexican Women in New Brunswick: Barriers and Opportunities

Miraida Morales, Ph.D. student, Xiaomu Zhou, Ph.D.
School of Communication and Information, Rutgers University, New Brunswick, NJ

Introduction
Half of New Brunswick residents are Hispanic, and yet not enough is known about this community (U.S. Census, 2010). A recent study showed significantly higher risk of problems of access to healthcare for Hispanic residents in New Brunswick (Stanley et al., 2008). Most of the existing research tends to lump Hispanic communities or Latinos into a homogenous group, resulting in a limited understanding of these communities. For instance, in the context of our study, many Mexican immigrants in New Brunswick, NJ, are of indigenous origins; grouping them with other Hispanic communities is not always judicious (Guarnaccia et al., 2012). Inspired by earlier studies, we seek to present a much more nuanced profile of the situation of this particular community in New Brunswick, in the context of their individual and community health practice. Due to women’s historical role in this community as primary caretakers for their families, they are the key for us to begin an exploration of the community’s health practices.

Objectives
The present study investigates (1) how Mexican immigrant women are able to continue to practice traditional or folk health practices when they settle in New Brunswick, and (2) how these folk health/medicine practices might be in conflict with health care programs, methods or treatment options they have access to in New Brunswick. Particular attention is paid to preventive health care methods, both as practiced by the community in their traditional or folk sense and as endorsed and promoted by the health care system in the United States, in order to identify when traditional methods clash or are modified vis-à-vis the dominant system as this group settles in the United States.

Method and Analysis
Several qualitative methods were used in this study including field observations, a structured survey questionnaire, and 2 focus groups. Theoretical sampling was conducted to identify study participants. In collaboration with community organizations, we conducted 2 focus groups with 14 Mexican immigrant women. Participants filled out a detailed demographic survey questionnaire. Focus group interviews were audio recorded to aid in transcription and analysis. Survey data was analyzed using descriptive statistics. Focus group qualitative data was transcribed and coded using a 2-step thematic analysis based on the Grounded Theory approach.

Preliminary Findings
Initial findings reveal several interesting themes related to the health practices of this community, including the role of food preparation and eating for maintaining health, how indigenous knowledge and cultural heritage influence members’ beliefs about health, and how members are able to maintain their health in their new environment (i.e. self-efficacy). Analysis uncovered the particular experience of rural and often indigenous immigrants in a large urban environment. Additionally, several institutional and systemic barriers emerged to community health practices, such as local agency policies, restrictions associated with the use of health insurance plans, local economic factors, interpersonal communication factors, and practices sanctioned by the local healthcare system.

Conclusion
The plurality of health practices in New Brunswick creates a complex environment that individuals must learn to effectively negotiate as they carry out the activities they associate with maintaining good health for themselves and their families. To better design and deliver health services to the Mexican immigrant community, we need to re-conceptualize health as a socio-cultural practice and to rethink the concept of “access” as these inform the context within which community health practices take place. Our research is still on-going, and we hope to further explore how culture and social factors determine health practices for the immigrant community, and how Health IT may help improve communication between the healthcare system and immigrant communities to support improved health outcomes through the development of socially and culturally relevant interventions. Health IT has the opportunity to address many of the communication and systemic barriers uncovered in the analysis.

References
Implementing Customizable Asthma Action Plans into an Electronic Medical Record

Stephen J Morgan, MD; Harley Z Ramelson, MD, MPH

1Partners HealthCare, Wellesley MA; 2Harvard Medical School, Boston, MA; 3Massachusetts General Hospital for Children, Boston MA; 4Brigham and Women’s Hospital, Boston MA

Abstract

An Asthma Action Plan is a written document that is provided to the patient in order to provide instructions on how to care for asthma symptoms at home and how to manage exacerbations. This poster describes the implementation of an electronic asthma action plan within a home-grown Electronic Medical Record and discusses some preliminary research data.

Introduction

The Longitudinal Medical Record (LMR) is an in-house developed ambulatory Electronic Medical Record used across Partners HealthCare. The LMR is a ONC-certified complete electronic health record. Included within the Patient Chart modules are actionable forms. The Asthma Action Plan (AAP) was first developed by the National Asthma Education and Prevention Program1-4 as a text based form that required manual entry of medications and specific instructions for the patient. This form is broken into three zones: green, yellow, and red, corresponding to the severity of respiratory symptoms. For example the green zone is a state of health with little or no symptoms while the red zone represents severe symptoms.

Features of the System

A text based AAP was initially developed in the LMR in 2007 and enhanced over time. The current AAP provides the ability to create the zones based on symptoms, predicted peak flow or personal best flow and to indicate the patient’s triggers. There is a direct link to the patients medication list, customizable favorite lists for each zone of the action plan, the ability to add of patient education forms and school based administration order forms, the ability to prescribe medications directly from the form, the ability to translate the form into Spanish and the ability to copy information from a previous AAP when creating a new AAP. Once completed, this form is made available in the Partners patient portal, Patient Gateway. An adult version of the form was recently released.

Results/Conclusions

Preliminary analysis of the usage data shows a steady increase in usage over the past 8 years. Overall, over 17,000 AAP forms have been created. As a result the future plans of this project are to assess the impact of adapting the form for adult patient use rather than primarily child health based. An additional area of research will be to investigate the impact on patient quality and improvement of asthma symptom control.

References

Extending the Project HealthDesign Experience via On-Line Public Data Repositories

Andrew J. Morland¹, Patricia Flatley Brennan¹, RN, PhD, FAAN
¹University of Wisconsin-Madison, Madison, WI

With the explosive growth in health tracking apps and the growing acceptance of their use in clinical care and self-management, app designers and developers are challenged to show validity and reliability of their tools. Indeed, the FDA Safety and Innovations Act calls for a risk-based regulatory approach to demonstrate app performance prior to their use in clinical practice. It is infeasible for every app developer to run a community-based clinical trial. Having access to data collected under similar circumstances would facilitate product testing and demonstrate app quality.

In Project HealthDesign five teams designed PHR apps that supported self-tracking of patient generated data. These teams generated thousands of data points representing time-stamped encounters used by patients to, for example, record mood or document meals. Grantee teams agreed in principle to deposit anonymized data generated by their apps into a public sharing repository. The Project HealthDesign National Program Office (NPO) selected ICPSR, the Interuniversity Consortium for Political and Social Science Research at the University of Michigan (www.icpsr.org).

ICPSR holds over 500,000 research data sets in the social sciences, including specialized data in education, aging and health among many other fields. ICPSR provides an efficient avenue for secondary data usage that allows researchers to bypass the onerous task of collecting primary data. The data sets that exist in the ICPSR repository are available to the public at no charge, and no registration or special credentials are required to view and use the data.

Creation, curation and use of data in research data repositories is more common in the bioinformatics research community than in the clinical informatics research community. Clinical informatics can contribute to and benefit from research data repositories. Primary investigations are expensive, and simply storing data after research questions are answered is wasteful. Research funding for individual investigators is becoming more challenging, necessitating investigators to seek new ways to answer research questions.

There have been notable efforts in the biomedical informatics communities to create large-scale data repositories. The Cancer Bioinformatics Grid (caBIG) espoused a bold vision to accelerate clinical trials and cancer research through a suite of data management, data storage and data sharing utilities. caBIG fell short of its intended goals and is defunct but the lessons about curation and data use policies remain valuable. NIH now maintains over 50 public repositories of research data. The Registry of Research Data Repositories (www.re3data.org) maintains an enumeration of over 1000 research data repositories and the APIs and related tools.

We prepared and deposited the Project HealthDesign data in ICPSR in Spring, 2015. The data pipeline is depicted in Figure 1. Each team collected the data according to their own research design and coding schemes. The NPO oversaw the curation of the data, which included over 12 months effort at standardizing variable labels, creating common code books and working with the individual teams to interpret data definitions. 13 data files including over 10,000 time stamped self-tracking events from over 100 unique patients were deposited.

The data deposited from Project HealthDesign has already been used in federally funded projects. Data from the BreathEasy project served as an initial data source for building a personalized prediction model.

Leveraging expensive field investigations through data sharing in public repositories can accelerate the development of high quality apps and contribute to patient centered care. (Support RWJF/NSF 1343969)

Figure 1. Data Collection, Curation and Use Pipeline (www.icpsr.org)
RECRUIT: Roadmap to Enhance Clinical Trial Recruitment Using Information Technology

Tasneem Motiwala, PhD; Chaitanya Shivade, BE; Satyajeet Raje, MS; Albert M. Lai, PhD; Philip R.O. Payne, PhD, FACMI
Biomedical Informatics, The Ohio State University, Columbus, OH

Introduction
Successful completion of randomized clinical trials, gold standards in medical research, is key to the progress of clinical and translational research. A major determinant of the success of a clinical trial is its ability to recruit the required number of participants. However, meeting the goals of trial recruitment is hindered by difficulties in identification of eligible patients that is, in part, due to complex nature of inclusion and exclusion criteria listed for trial protocols as well as the inability to efficiently match the criteria with patient’s clinical data. Despite efforts to alleviate obstacles associated with screening for eligible participants, statistics on clinical trial participation remain static. The purpose of this study is to automate the screening process through utilization of electronic tools coupled with a decision tree approach to sequentially narrow the search space in terms of all patients eligible for a specific clinical trial or vice versa.

Methods
Due to availability of clinical data for patients diagnosed with chronic lymphocytic leukemia (CLL), we will use CLL as a disease domain to test the feasibility of our proposed model. Records for all CLL trials conducted at The Ohio State University and listed in clinicaltrials.gov were downloaded as XML files. Eligibility criteria extracted from these records were segmented into individual sentences using the Stanford NLP tokenizer. Conceptual graphs were then used to identify similar concepts within the individual sentences (Fig. 1A) to determine their frequency of occurrence. Among these concepts, criteria that would be potentially useful in screening with the help of electronic medical records (EMRs) were identified and data corresponding to these criteria were classified as being present in discrete fields or in notes within the EMR.

Results and Discussion
In all, 84 relevant trials were identified from clinicaltrials.org. Some of the frequently occurring concepts for which data would be present in discrete fields within the EMR include pregnancy/nursing status, hemoglobin levels, HIV infection status, etc (Fig. 1B). Other frequently occurring concepts requiring extraction of information from notes include ECOG status (staging of CLL), presence of organomegaly, and fatigue levels (Fig. 1C). Some of the less frequent concepts were neutrophil counts, congestive heart failure, bilirubin levels, etc (Fig. 1D). Next steps would be to determine the populace of data within the discrete fields and the efficiency of extracting the necessary information from clinical notes. Based on these, a finalized set of useful criteria, including rare and frequently occurring ones, will be compiled and a CLL data mart will be created to hold data corresponding to these criteria. Algorithms will be developed to prioritize and order the eligibility criteria such that at every step the search space i.e. number of potentially eligible patients will be reduced. Data within the CLL data mart will be used to test these algorithms.

Figure 1: Example of a conceptual graph (A) and frequencies of frequent/discrete (B), frequent/notes (C) and infrequent (C) concepts.
An automated tool to replicate data between multiple versions of Profiles Research Networking Software’s (RNS)

Sukrit Mukherjee, MS, MSE, PhD¹, Sebastien Delta, BS¹
¹Charles R. Drew University of Medicine and Science, Los Angeles, CA, USA

Abstract

At Charles R. Drew University of Medicine and Science (CDU), we adopted different versions of Profiles RNS to use locally and to feed their data to other research consortiums. In order to synchronize information between different instances of Profile RNS which are architecturally different, we created an automated tool to implement replication of data involving complex cross-referencing, number of inner and outer joins and Node table lookup across different versions of Profiles RNS databases.

Introduction

At CDU, researchers are involved with clinical and translational research, especially focusing on health disparities. Center for Biomedical informatics provides investigators at CDU with state-of-the-art informatics solutions for their research projects. In 2011 We had adopted and heavily customized NIH-funded open source tool Profiles Research Networking Software (RNS) Beta Version and called it
Variation in EHR Implementations and the Impact on Safety of Test Result Follow-up

Daniel R. Murphy MD MBA\textsuperscript{1,2}, Michael W. Smith PhD\textsuperscript{1}, Dean F. Sittig PhD\textsuperscript{3}, Elise Russo MPH\textsuperscript{1}, Hardeep Singh MD MPH\textsuperscript{1,2}

\textsuperscript{1}Houston VA HSR&D Center for Innovations in Quality, Effectiveness and Safety, Houston, TX.
\textsuperscript{2}Baylor College of Medicine, Houston, TX.
\textsuperscript{3}University of Texas Health Science Center, Houston, TX.

Background
Breakdowns in the process of electronic health record (EHR)-based test result ordering and follow up are common and can negatively impact patient care. The causes of such breakdowns are multifactorial and include both organizational/social issues (e.g., workflow, training) and technical factors (e.g., EHR interface design). Increased understanding of these factors is needed in order to make test results follow-up and tracking fail-safe. As part of ongoing work to understand socio-technical factors involved in breakdowns of test results follow up, we sought to identify differences in EHR implementations that impact test result follow-up processes.

Methods
Employing an 8-dimension socio-technical framework for EHR implementation and use, we conducted literature reviews and gathered expert opinion from a multidisciplinary team to identify factors that could either positively or negatively impact fail-safe test result tracking and follow-up. We recruited 3 large healthcare facilities using commercial EHRs and one Veterans Affairs (VA) facility and performed EHR demonstrations at each site. For consistency and completeness, we developed and used an observation checklist and set of interview questions. Demonstrations were audio recorded, and detailed notes on each step were taken. The demonstrator (training staff) was asked to perform specific tasks related to the process of test ordering and result tracking and follow-up. For each step in the tasks, we ascertained whether certain factors were present and positively or negatively impacted fail-safe tracking and follow-up. Any between-site differences were noted, and the multidisciplinary team reviewed all differences to determine whether they provided better, worse, or the same level of safety as the features that existed at all other sites. A second round of demonstrations was performed to address any questions regarding survey items that emerged after the initial demonstrations.

Results
A total of 8 demonstrations were performed at each of 4 large healthcare facilities. Evaluations included two sites with Epic Hyperspace, one with GE Centricity, and one with the VA’s Computerized Patient Record System (CPRS). We found at least one ‘positive’ factor at each site not present at other sites, and only one site without any unique ‘negative’ factors. Positive factors* not present at all sites included: ability to track orders to completion (2 sites); ability to monitor results unopened by providers; use of structured data to code abnormal radiology reports; use of “Addended Alerts” to inform providers when original results were amended/modified; presence of useful contextual clinical information in the test result review screen (e.g., next visit date); on-site information technology (IT) staff support; and ability to send reminders to self at a future date (2 sites). Negative factors* not present at all sites included: presence of orderable items that do not lead to action (listed as “Do Not Order”); automatic removal of test results from the inbox after a pre-set time-period even if not reviewed by the provider; inability to determine if result is abnormal from the result review screen; laboratory result graphs that were difficult to interpret (2 sites); and inability to return a result to the inbox after it is removed without IT assistance (e.g., if a provider is interrupted during processing).

Conclusions
Real-world EHR implementations are accompanied by several factors that positively or negatively impact the safety of test results follow-up and tracking processes. Understanding of these factors can allow for the development of best practices for EHR design, implementation, and use that ensures safer tracking and follow-up of abnormal test results.

*Positive and Negative factors were identified at a single site, except were noted.
HDD Access – an Open Source Terminology Server with Publicly Available Healthcare Terminology Content

Senthil K. Nachimuthu, MD, PhD; Lee Min Lau, MD, PhD
3M Health Information Systems, Inc., Murray, Utah

Abstract

HDD Access is a publicly available terminology server published since 2012 by 3M with support from the Department of Defense (DoD) and the Department of Veterans Affairs (VA). HDD Access supports easy implementation, maintenance and use of both standard and local terminologies. It includes standard terminologies such as RxNorm, ICD-9-CM, ICD-10-CM, ICD-10-PCS and others. It also allows users to add their own terminology content. We will demonstrate the HDD Access terminology server, its terminology content, browsing, searching and authoring tools, web services API, and the lessons learned in creating a publicly available open source product. We will demonstrate how standard terminologies can be implemented easily, how users can create local extensions to support their own terminologies, and how applications can use terminologies through a web-services API based on HL7 Common Terminology Services.

Description of the System

HDD Access is a concept-based terminology server with content covering various domains within healthcare, which serves as an interlingua that spans the needs of several healthcare applications. HDD Access supports local extensions, permitting the users to integrate their local terminologies into their instance of HDD Access, which enables interoperability with standard terminologies. HDD Access enables semantic interoperability and terminology governance by following essential principles of terminology design, such as concept orientation, concept permanence, nonsemantic concept identifier, polyhierarchy, multiple granularities, multiple consistent views, graceful evolution, recognition of redundancy, and a framework for formal definitions, described in Cimino’s Desiderata (1998), and enables users to easily implement these best practices.

Innovative Features of the System

HDD Access is unique among publicly available free-of-cost terminology servers by including a concept-based terminology with about 500,000 concepts at present. HDD Access software is written in the Java programming language, supports multiple open source and proprietary relational databases (such as MySQL, PostgreSQL, Oracle, Microsoft SQL Server and H2) and works on both Windows and Linux operating systems.

HDD Access includes user-friendly web-based tools for browsing and authoring terminology content. It also includes an advanced search engine which helps to search concepts by finding synonyms, inflections and other variations of the input terms. HDD Access includes a RESTful web services API that supports both HL7 CTS v1.2 and 3M functions for querying and authoring the content.

HDD Access allows the users to load and integrate their own local terminologies as well as map them to standard terminologies while maintaining control over their content through a novel namespace design, which automatically manages inter-dependencies between various terminologies so that they can be authored, imported and exported while preserving content integrity.

Current Implementation Status

HDD Access was first released in August 2012 and monthly updates are published through the HDD Access website, located at www.hddaccess.com. The HDD Access terminology content and software are downloaded about 150 times per month on average. The HDD Access user community has more than 1,500 users as of February 2015. The entire HDD Access software package is distributed as open source under the Apache Software License version 2, which allows modifications and derivative works, including proprietary applications by end-users. The entire HDD Access terminology content is distributed free of cost under the HDD Access Content License, which permits commercial use as well as local extensions.
Large Scale Regional Medical Information Network Infrastructure

Naoki Nakamura, Masaharu Nakayama, Takuo Suganuma, Norio Shiratori, Teiji Tominaga
Tohoku University, Japan

Abstract
Since the Great East Japan Earthquake, we have been developing regional medical information network systems. The private cloud and network technologies were introduced to realize their pliability and scalability. In this paper, we describe network infrastructure essential for a large scale regional medical information network system.

Introduction
With the budget of the Japanese Ministry of Internal Affairs and Communications grant, we have been developing EHR (Electronic Health Records) systems named, Miyagi Medical and Welfare Information Network (MMWIN), where institutions, such as hospitals, clinics, pharmacy, and nursing homes upload medical treatment and health information records of patients and residents into the centralized storages. To realize a large scale EHR system over the whole prefecture, we built an advanced network infrastructure with pliability and scalability for this MMWIN system.

Large Scale Network Architecture for EHR
In MMWIN, various efforts are carried out to keep the system secure and stable at a lower cost. Figure 1 shows the overview of the network where firewalls are placed according to the security levels. Servers and switches, firewalls, load balancers, and routers are virtualized. Networks having different security levels are separated from each other by means of VRF (Virtual Routing Forwarding) technologies on both data center and branches. There are multiple virtualized routers and firewall on both data center and branches according to the security level so that each can be separated completely. As the tele-conference’s traffic passed through the internet, it can be completely separated from other traffic. In addition, VRF enables routers to run multiple routing instances, then it handles multiple network easily and securely as if there are multiple routers. DMVPN enables direct tunneling among branch’s IPsec routers. When a spoke router wants to transmit a packet to another spoke router, these two routers dynamically create an IPsec tunnel, and traffic is directly transferred to each other. With this technology, tele-conference traffic via data center can be diminished and as well as configuration of routers becomes simpler. System keeps high redundancy in both the access lines and network equipment becomes redundant, and that routing is controlled with dynamic routing, EIGRP. These technologies are essential for a large scale regional medical information network system in terms of pliability and scalability.

Evaluation
Unlike other EHR systems, MMWIN can not only provide multiple networks having different security levels, but also can cable of virtually closing secure networks. In addition, it is considered to be redundant configuration. To realize these networks according to different security levels, usually systems are very complicated. Although MMWIN can attain the simplification of system configuration effectively with advanced network technologies, such as DMVPN, VRF, and Dynamic Routing. Regarding scalability, we had verified the capacity, that this network infrastructure can run well with more than 700 routers in advance. Currently there are more than 450 branches of this network, problems such as congestion have not occurred and system can be kept stable.

Conclusion
We discussed about a network infrastructure for a large scale regional medical information network system. For more flexible operations, we are planning to develop management tools to provide services in a more effective manner.

Figure 1. Overview of MMWIN Network Constitution
A Prefectural Medical Information Network System Developed after the Great East Japan Earthquake
Masaharu Nakayama1,2, Hiroaki Shimizu3, Kazuo Sato4
Naoki Nakamura2; MMWIN investigators.

1Disaster Medical Informatics, International Research Institute of Disaster Science, Tohoku University, Sendai, Japan; 2 Medical Informatics Center, Tohoku University Hospital, Sendai, Japan; 3 Neurosurgery, Akita University Hospital, Akita, Japan; 4 Yamoto Clinic, Higashi-matsushima, Japan

Abstract
A medical network system, the Miyagi Medical and Welfare Information Network (MMWIN), has been in development since 2013. MMWIN allows the backup of medical information of hospitals, clinics, pharmacies, and care facilities. By the end of February 2015, more than 1.6 million registered patients and more than 42 million pieces of backed up data were recorded on it. Increasing patient registration and maintaining a balance between income and cost are critical to maintaining this project.

Introduction
The Great East Japan Earthquake of March 11, 2011, which devastated the northeastern coastal region of Japan and resulted in 15,890 deaths and 2,590 missing persons as of March 7, 2015, is one of the largest disasters on record. Many medical facilities were destroyed and medical information, whether on paper or on servers in hospitals and clinics, was lost. A shortage of medical information disrupted clinical activity. Japan recognized the importance of backing up patient data after the disaster.

MMWIN
In Miyagi prefecture, one of the areas most damaged by the Great East Japan Earthquake, a project to develop a network system among hospitals, clinics, pharmacies, and care facilities was launched in 2013. The system is the Miyagi Medical and Welfare Information Network (MMWIN). The government provided funding to the network throughout the prefecture for three years. The medical information of each facility has been backed up in case of a disaster. Moreover, the information has been shared among medical facilities, with the consent of registered patients. The backup system is based on the Standardized Structured Medical Information eXchange (SS-MIX), which enables data from medical record systems developed by different vendors to be stored in a similar format. As of February 24, 2015, the number of facilities with the MMWIN was 390. The number of registered patients was 1,697,850, or more than half the residents of Miyagi prefecture. Medical data—including patients’ basic information, disease names, blood tests, and prescription lists—included 42,578,989 items.

Problems
The MMWIN has encountered several problems. First, development has been dependent on the budget. Most income has come from the fees of medical facilities. However, the gap between income and the cost to maintain the system has remained large. Thus, sustaining the project is a great burden. Second, image data are not available, a shortcoming that distracts clinicians. Third, access to detailed clinical information is limited by the content that is required.

Conclusion
The MMWIN started in one area—Miyagi prefecture—after the Great East Japan Earthquake and has become prevalent throughout the prefecture. The number of registered patients and clinical data are increasing. However, several problems must be solved to maintain the system in the long term.

References
Evaluation of Perioperative Medication Errors and Adverse Drug Events

Karen C. Nanji, MD, MPH1,2,3, Sofia Shaikh1, Amit Patel, MD1, Diane L. Seger, RPh3,4, David W. Bates, MD, MSc2,3,4
1Massachusetts General Hospital, Boston, MA, Country, USA; 2Harvard Medical School, Boston, MA, USA; 3Partners Healthcare Systems Inc., Wellesley, MA, USA; 4Brigham and Women’s Hospital, Boston, MA, USA

Introduction: While perioperative medication administration presents particular patient safety challenges compared to other hospital settings, the literature on perioperative medication errors (MEs) is sparse and contains largely self-reported data.1,2 The validity and reliability of studies based on self-reporting of MEs in other patient care areas has been called into question.3,4 The purpose of this study is to assess the rates of perioperative MEs and adverse drug events (ADEs) in the perioperative setting, evaluate their root causes, and formulate targeted solutions to prevent them.

Methods: We iteratively revised a previously validated error detection framework to define what constitutes a ME and/or ADE. Anesthesia-trained study staff subsequently observed randomly selected surgical cases at a large tertiary academic institution to identify MEs and ADEs over a period of seven months beginning in November 2013. Chart abstraction was also performed to flag MEs and ADEs that may have been missed by observation. All events identified subsequently underwent independent review by two specially trained reviewers. Our primary outcomes were the incidence of MEs and ADEs in the perioperative setting.

Results: A total of 277 operations were observed with 3,671 unique medication administrations of which 193 (5.3%, 95% CI 4.5-6.0) involved a ME and/or ADE. Of the events identified 153 (79.3%) were MEs in which 32 had little potential for harm, 70 had potential to cause harm (near misses) and 51 resulted in patient harm (preventable ADE). Of the 91 (47.2%) identified ADEs, 40 (20.7%) did not involve a ME and 51 (26.4%) were due to MEs that resulted in patient harm. Of the 153 observed medication errors, 99 (64.7%) were serious in severity, 51 (33.3%) were significant and 3 (2.0%) were life threatening. The most common error types were labeling errors 37 (24.2%), wrong dosage errors 35 (22.9%) and omitted medications/failure to act 27 (17.6%). Medications most frequently associated with errors were propofol (30, 25.6%), phenylephrine (12, 10.3%) and fentanyl (11, 9.4%). Longer procedures, especially those greater than six hours, had higher total event rates (p = 0.000), ME rates (p = 0.000) and ADE rates (p=0.004) than shorter procedures. Also, procedures with 13 or more medication administrations had higher event rates (p=0.02) and ADE rates (p=0.002) than those with 12 or fewer medication administrations.

Discussion: We found that approximately one in twenty perioperative medication administrations included an ME and/or ADE. More than one third of these errors led to observed patient harm, and the remaining two thirds had the potential for patient harm. These rates are markedly higher than those reported by existing retrospective surveys in the literature.1 Based on the error types we found, we identified several strategies to minimize perioperative MEs and ADEs. These strategies include both technology-based interventions and process-based interventions. Examples of technology-based interventions include bar code-assisted syringe labeling systems, point-of-care bar code-assisted anesthesia documentation systems, specific drug decision support, and alerts. Process-based interventions include determining optimal timing for documentation, reducing opportunities for workarounds, connecting infusions to the most proximal IV port, rigorous vendor selection and strong training. Future analyses should target the implementation of process- and technology-based solutions that address the root causes of these errors in order to reduce their incidence.

References

Automated Approach to Extract Cardiovascular Phenotypes from Echocardiography Reports

Chinmoy Nath, PhD¹, Mazen S. Albaghdadi, MD², Siddhartha R. Jonnalagadda, PhD¹
¹Department of Preventive Medicine, ²Center for Cardiovascular Innovation, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Abstract
We developed an information extraction system, EchoInfer, using natural language processing (NLP) that extracts 66 cardiac structural and functional data elements from echocardiography report. Our automated approach to the characterization of cardiovascular phenotypes from reports containing structured, semi-structured, and unstructured data has important implications for the development of research studies and has the potential to impact patient care. EchoInfer is unique because of its comprehensive coverage of phenotypes present in echocardiography reports and an overall high accuracy of over 90%.

Introduction and Background: The goal of this study is to explore the feasibility and reliability of using natural language processing (NLP) for the extraction of multiple data elements from echocardiography reports. Echocardiography is one of the most commonly ordered diagnostic tests in cardiology. However, tools for the effective extraction of information from echocardiography reports that may inform clinical management and analysis of data for research purposes are lacking.

Participants and Methods: Echocardiography reports from 1547 patients (1053 men and 494 women, mean age 67.9±14.0 years) with a history of aortic bioprosthetic valve implantation were considered. Echocardiography reports contained four different sections: (i) report text, (ii) procedure components, (iii) measurement and calculations, and (iv) conclusion or summary. Sentences were separated from each other in echocardiogram reports and the sentences that do not contain any data element were discarded. We developed an array of regular expressions and rules based on initial set of data to extract 66 cardiac structural and functional data elements such as left ventricular ejection fraction (LVEF), valvular gradients and velocities from all sections of the report. A training set consisting of 5 batches of 20 randomly selected echocardiography reports were used to further modify the regular expression patterns. The output of the NLP analysis was then presented in a structured format that was amenable to analysis by the end-user for possible research and/or clinical purposes. The accuracy of EchoInfer was evaluated by an independent reviewer using another randomly obtained test set of fifty echocardiogram reports.

Results: The accuracy of EchoInfer was evaluated by an independent reviewer using another randomly obtained test set of fifty echocardiogram reports; i.e. 50*66 = 3300 phenotype values. EchoInfer was evaluated with widely accepted definitions of true positives, false positives and false negatives [1]; and it achieved a sensitivity (recall) of 94.03%, a positive predictive value (precision) of 95.64% and an F1-score of 94.83%.

Discussion and Conclusion: Previous attempts to characterize cardiovascular phenotypes using NLP from echocardiography reports have been limited by characterization of only a few cardiac data elements, such as LVEF [2]. We showed that NLP provides a powerful and reliable tool for the large-scale, comprehensive extraction of information from unstructured clinical notes such as echocardiography reports. The use of EchoInfer may have implications for the clinical management and research analysis of patients undergoing echocardiographic evaluation.

Acknowledgements: This work was partly funded by National Library of Medicine grant R00LM011389.

References

1611
Development of a Methodological Protocol for Observing Pharmacist Information Needs While Using the EHR

Scott D. Nelson, PharmD, MS; Joanne LaFleur, PharmD, MPH; Guilherme Del Fiol, MD, PhD; R. Scott Evans, MS, PhD, FACMI; Charlene R. Weir, RN, PhD, FACMI

Introduction
The adoption of electronic health records (EHRs) has impacted how physicians, nurses, and pharmacists provide care for patients. Although many studies have examined EHR use by physicians and nurses, no studies have examined pharmacists' information needs or information seeking behavior in the context of EHR use. The Veterans Health Administration (VHA) is an ideal setting for studying pharmacists' information needs, as there is a long-established EHR as well as advanced clinical practice for pharmacists. However, standard methods of assessing information needs impose substantial challenges in this setting, including significant recall bias in interviews and surveys; the need for multi-media capture due to the rapid, dense, and variable nature of data in real clinical settings; restrictions on installing software (such as screen capture) in many institutions; and typical unavailability of access to log data along with the inability of log data to capture the use of external information sources (such as paper or whiteboards). Additionally, think-aloud techniques conducted in real life settings are very limited as experts have difficulty verbalizing their goals and tasks; and most people stop talking when they are thinking or processing complex information. Historically, eye-tracking tools have had variable success in the real clinical work settings due to their obtrusiveness and the granularity of data. To address these challenges, we developed and piloted a flexible method that could be used in other clinical settings, could capture gaze, and could minimize interruptions. The purpose of this poster is to describe the information integration benefits and challenges to this approach.

Methods
We observed seven clinical pharmacists in the inpatient setting while they reviewed patients in preparation for medical team rounds. The pharmacists were asked to prepare for rounds as they normally would in their work environment, evaluating real patients, and interacting with the medical team. To record the observation sessions and pharmacists' interactions, we used an eye-tracking camera from Pupil-Labs. During the observation, pharmacists were asked clarifying questions to better understand goals, mental models, responsibilities, perceived usefulness of different sources of information, and tasks. We would wait to ask the questions until subjects completed a task as to not disrupt the cognitive process of complex tasks. A researcher would also document field notes during the session and ask deepening questions about information needs and goals after the session. We used an iPad to record audio from the session that was merged with the video using Adobe Premiere prior to analysis.

Results
Overall, the study method was well received. Pharmacists were extremely supportive and excited to participate and use the eye-tracking device. There were no complaints of disruption, interference, or discomfort due to the device or study procedure, although it seemed somewhat distracting to other clinicians. Some of the benefits include the low cost of equipment; augmentation of observation and interviews with video-audio recording and probing questions; and immersion of the researcher in the pharmacists' environment with little disruption to the normal tasks. Some of the challenges included the lack of native audio recording with the Pupil-labs camera, video codec difficulties, and problems correctly correlating the pupil recordings with the actual location of the pharmacist's gaze. The video codec used by Pupil-Labs software (Matroska) is not natively supported on MacOS, so the video file had to be converted before importing into Adobe Premiere. While the pupil tracker did not provide the precise location of the pharmacist's gaze, it did provide a general area of gaze that helped in review of the verbal and observational data.

Conclusion
Using an eye-tracking camera along with contextual inquiry helped embed the researcher in the pharmacists' environment and overcome many challenges of studying VHA pharmacists' information needs in the EHR.

References
Early Detection of Heart Failure using Data Driven Modeling Approaches on Electronic Health Records: How far can one go without Domain Knowledge?

Kenney Ng, PhD1, Jianying Hu, PhD1, Yajuan Wang, PhD1, Walter F. Stewart, PhD2, Steven R. Steinhubl, MD3, Christopher deFilippi, MD4

1IBM Research, T.J. Watson Research Center, Yorktown Heights, NY; 2Sutter Health Research, Walnut Creek, CA; 3Geisinger Health System, Danville, PA and Scripps Health, San Diego, CA; 4Univ. of Maryland School of Medicine, Baltimore, MD

Abstract
This retrospective study on a 15K patient cohort investigated the feasibility of developing effective heart failure onset risk prediction models by applying data driven modeling approaches to electronic health record data. A predictive model built without using any domain knowledge about the data achieved an AUC of 0.802 (95% CI: 0.790-0.814).

Introduction
Heart failure (HF) is a major public health issue with increasing prevalence worldwide. Early recognition of HF pathophysiology prior to a clinical diagnosis would provide the means to implement lifestyle modification and pharmacotherapy interventions that may slow disease progression and improve patient outcomes. Developing knowledge driven models for disease onset risk prediction can take a significant amount of time and effort with data curation, data understanding, information extraction and modeling. It also requires domain specific expertise in all phases of the process. With the increasing amount of electronic health record (EHR) data available and advances in machine learning and data mining techniques, we examined how the use of generic data driven modeling techniques could be used to develop predictive models for incident onset of HF from EHR data without domain knowledge. The expectation is that such a data driven model can be used initially and later supported by domain specific knowledge to increase performance, improve interpretability and refine the clinical applicability of the model.

Methods
The study subjects consisted of 15,209 patients: 1,684 incident HF cases and 13,525 gender, age and physician matched controls. The longitudinal EHR contained demographic, vital, diagnosis, lab and medication data totaling over 32K unique variables. The creation of the case and control patient cohorts required domain knowledge, but the subsequent modeling did not use any domain specific knowledge about the patient data, i.e., the type and meaning of each variable was unknown. A feature vector representation for each patient was constructed based on the EHR data by specifying an observation window (two year duration, one year before the diagnosis date) and then aggregating all instances of the same variable within the window into a single value. Basic aggregation methods (Boolean, count, sum and mean) were used. Feature selection using the information gain measure was used to identify the top N features and predictive models (random forest classifiers with 100 trees) were trained using 10-fold cross validation. Performance was reported using area under the ROC curve (AUC) and 95% confidence interval (CI) metrics.

Results
Figure 1 shows prediction performance (in AUC) as a function of the number of selected features for five different feature construction aggregation methods: sum, count, Boolean, mean, and a type specific approach where mean was used for continuous numeric variables (labs and vitals) and Boolean was used for categorical variables (demographics, diagnoses and medications). For all aggregation methods, performance improved as more features were added (up to 500 variables) and then leveled off. The type specific aggregation method (●) performed better than the other methods and had an AUC of 0.802, 95% CI: 0.790-0.814 with 500 feature variables.

Discussion and Conclusion
Given a pre-defined set of case and control cohorts, an effective HF onset risk prediction model was developed by applying generic data driven modeling approaches to EHR data without using any domain knowledge about the data. As next steps, we want to quantify how much value (in terms of prediction performance) is added by incorporating successively more domain knowledge into the modeling process with a goal of trying to identify where use of domain knowledge would be the most productive in developing disease risk prediction models.
Role of Social Media in shaping public health messages regarding Human Papillomavirus (HPV) vaccinations

Lien B. Nguyen, MPH\(^1\), Priya Nambisan, PhD\(^2\)
\(^1\)College of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI; \(^2\)College of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI

**Introduction**

Human Papillomavirus (HPV) is the main cause of cervical cancer. Cervical cancer affects an estimated 14,000 women each year and 4,000 deaths annually in the United States (Center for Disease Control and Prevention, 2015). HPV is a common virus that is passed through sexual contact between persons. Cervical cancer is a preventable disease, when found early, through proper detection of health screenings and also through HPV vaccinations. The Internet, especially social media, has become a major platform for individuals to quickly access and share important information about health concerns. Social media sites such as Facebook and Twitter are utilized by millions of people each day for this. According to the Pew Research Center, a survey reported 81% of U.S. adults use the Internet and 72% say they have looked online for health information during the past year (2013). Information about cervical cancer and HPV vaccinations can be found, shared, and posted by different organizations and peers among popular social networking platforms. However, current research studies that are focused specifically on social media and cervical cancer prevention are limited and we do not know how social media influences the shaping of overall message regarding HPV vaccinations. The consistency of different messages posted every day on social media sites, which are accessible by massive groups of users, may impact the individual’s decision to take important steps towards preventative measures, such as to obtain HPV vaccinations, in order to help fight against cervical cancer.

**Objective** – The study objective is to determine how social media play a role in shaping public health messages especially in the case of HPV vaccinations.

**Methods** – The methods for this study included data collection from social media platform, Twitter, where we examined individual posts and comments discussing topics related to HPV vaccinations. Semantria program was used to perform Sentiment Analysis on data collected for a total of 4751 posts and comments, which ranged from June 2014 to May 2015. We classified them into positive and negative messages on HPV vaccinations. We further analyzed types of information and different sentiments shared among users about obtaining vaccinations for prevention against cervical cancer.

**Results** – This is a currently on-going research. Findings are expected to show how negative messages compete with positive ones in critical public health messages such as for HPV vaccinations. Detailed findings will be presented at the conference.

**Conclusion / Implications** – The study findings will reveal how current social media is used as a tool for communicating key information about HPV vaccinations to the online audience. From this, we will have an understanding of what types of actual information is being shared. The research findings will have important implications for future public health workers, online communications, and use of informatics to address public health problems.

**References**

Finding Similar Drug Classes using RxClass

Thang Nguyen, M.S., Lee Peters, M.S. and Olivier Bodenreider, M.D., PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Contact information: RXNAVINFO@LIST.NIH.GOV

Motivation
Drug classes constitute important information about the drugs and are critical to important use cases, such as clinical decision support (e.g., for allergy checking). Finding similar drug classes helps to identify important characteristics of drugs in different classification systems. RxClass, a web-based browser for drug classes, supports navigation between RxNorm drugs and drug classes from several sources (ATC, MeSH, DailyMed and NDF-RT), and allows users to explore similar classes.

Finding Similar Classes by Drug Membership
The RxClass application and its companion RxClass API offer functionality which allows users to retrieve similar classes relative to the selected class. Similar classes are defined here as classes which share a large proportion of drug members. The simple procedure for comparing two classes includes the following steps:

- Select a drug class (e.g., CALCIUM CHANNEL BLOCKERS (C08) in ATC) and display its members.
- Select the “similar classes” link above the list of drug members. A popup menu is displayed with the most similar classes (e.g., Calcium Channel Interactions in DailyMed). The class information along with the equivalence score is displayed for each similar class (here: 0.74).
- Select the “Venn” link to visualize shared vs. specific clinically significant drugs (ignoring base/salt distinctions) for the two classes as a Venn diagram, as well as a table. See figure 1.

In addition to retrieving similar classes for a given class, the RxClass API also supports the identification of classes from an arbitrary list of drug members (e.g., a drug value set).

Similarity Scoring
RxClass and the RxClass API similarity functions use equivalence and inclusion scores defined in prior work to find and rank similar classes. The equivalence score is based on the Jaccard coefficient, which we evaluated in previous work. The inclusion score identifies situations where a high similarity value is indicative of one class being included in the other.

Conclusions
Providing similar drug classes has been a missing piece of information in RxNav. With RxClass, we now provide a link between similar drug classes from various sources using its RxNorm drug membership.

RxClass is available from the main RxNav website (http://rxnav.nlm.nih.gov), along with additional information about RxNav and our APIs to various drug information sources.

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Figure 1. Sample screenshot of RxClass Similar Class Information
Multi-Agent (Team) Microworld Environments for Healthcare

Vickie Nguyen, MA, MS and Amy Franklin, PhD
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

Introduction

Much of healthcare relies on distributed clinical providers working together as a team to complete coordinated patient care tasks in complex clinical environments. However, these environments are not solely human based, but are socio-technical environments requiring interactions of humans with humans and human teams with technology systems. To create effective technology systems for clinicians to provide safe and high quality patient care, we believe it is vital to also design these systems according to team needs. Yet, how can we best evaluate how well these team-designed health information technology (HIT) systems actually fit team needs?

Creating a simulation of a multi-agent microworld may be the answer to effectively evaluating team HIT systems. To our knowledge, current healthcare team simulations mostly focus on examining training, levels of expertise, or individual technology system evaluations. This project is still in progress, but in this work, we briefly describe the creation of a multi-agent microworld of a complex clinical environment, the Emergency Department (ED), for the evaluation of team HITs and the microworld’s initial uses.

Methods

Microworlds are scaled-down simulations of the team task environment, are conducive for evaluating multiple participants simultaneously, have been utilized in other complex environment domains like aviation, and are created using cognitive engineering methods.

We based our ED microworld on a model derived from 226 hours of ethnographic non-participant observations and surveys of 485 Gulf Coast Trauma 1 Teaching Hospital ED clinicians. We focused on Attendings, Residents, Mid-Levels, and Nurses as they worked together to care for all patient types upon the patients’ entrance into the ED to their discharge or transfer from the ED.

From our findings, we created an ED environment where a team of clinicians must participate in multiple concurrent interactions with one another and a team HIT system to complete multiple coordinated patient care tasks. Figure 1 shows an example screenshot of our computerized ED microworld and its human to human and human to technology interactions.

We are currently in the process of evaluating how well a mobile application, designed to help facilitate ED clinical team communications, actually helps ED clinicians better manage interruptions in the ED. The ED microworld helps us simulate the necessary testing interactions.

Conclusion

Although still a work-in-progress, the multi-agent ED microworld seems promising in replicating real world interactivity and multitasking. Multi-agent microworlds can provide a helpful means for evaluating team HIT systems, and in turn, help create more effective technology systems for complex clinical environments.

References

Leveraging an Open Source Data Warehousing and Analytics Tool to Promote Longitudinal Research, Improve Knowledge Transfer and Avoid Redundancy Across Research Studies

Thomas Nutman, M.D.1; Alex Rosenthal, M.S.,M.B.A.1; Joe Croghan, M.S.1; Cathy Steel, Ph.D.1; Karla Feghali2; David Hardison, Ph.D.2

1NIH NIAID, Bethesda, MD; 2ConvergeHEALTH by Deloitte, Newton, MA

Abstract - A customized open-source platform designed to streamline research efforts for the NIAID / Laboratory of Parasitic Diseases (LPD) integrates data from multiple disparate data sources collected since 1974. Standardized data are organized around research subjects enabling scientists to perform retrospective analyses and plan prospective research studies. A customized individual patient profile provides a holistic and longitudinal view of research participants. Overall researchers experienced increased efficiency, accuracy, and cost-effectiveness of patient data management and analysis.

Introduction - The National Institute of Allergy and Infectious Diseases (NIAID) conducts and supports basic and applied research to better understand, treat, and ultimately prevent infectious, immunologic, and allergic diseases. The Institute manages and retains data collected from clinical and population health studies. As data collection and storage technologies evolved over time, data sets of different formats and structures have been dispersed across multiple sources including handwritten laboratory notebooks, Excel spreadsheets, and the NIH Biomedical Translational Research Information System (BTRIS).

Problem - To preserve existing clinical and scientific data for record keeping and potential re-evaluation and reproducibility, NIAID sought a flexible and extensible solution to store, maintain and extract diverse data sets from one centralized repository. Specifically, the Laboratory of Parasitic Diseases (LPD) was seeking a long-term solution to consolidate longitudinal data sets collected since the 1970s to avoid redundancy in re-running laboratory experiments and make any collected data available to current researchers and staff. Additionally, the integrity of data assets would be preserved for use in future research in the event that any scientist left NIAID, and thus, increase the ROI on the research funding that supported the initial data capture.

Solution - A customized version of the tranSMART version 1.2 Knowledge Management Platform, an open-source tool, was implemented that can integrate data from various clinical and field studies in a single standardized repository. This platform facilitates self-service analytics and is designed to help scientists develop and refine research hypotheses through a very user-friendly drag and drop interface.

The data in tranSMART is stored in a custom ontology or data hierarchy which was established in collaboration with researchers. This ontology allows users to quickly filter out their data of interest and perform various analyses by simply dragging and dropping patient cohorts into the interface provided by the tool. In addition, researchers are able to run rapid statistical analyses on data sets to support retrospective and prospective studies. An individual patient profile view was created which provides a temporal view of accumulated data, e.g., blood platelets assayed at multiple time points in a study of filarial infections. An incremental load feature allows researchers to upload new data to already existing patient records. This enables scientists to integrate their real-time findings with existing data. The tool also supports easy data selection and export enabling scientists to analyze their data using other analytical tools beyond the R-based selections in tranSMART. Both the data selection and analysis steps can be saved to a researcher’s personal workspace in tranSMART to facilitate reproducibility and collaboration.

Conclusion - Through data curation and loading into a central repository that is accessible through a self-service, easy to use interface, NIAID researchers can spend more time on scientific insight. Similarly, the Institute can achieve better returns on the funding applied to data capture, data management, and analysis while preserving its rich data assets for future use.
Development of Anatomical Radiology Ordersets

Kevin O’Bryan, MD, Charles H Andrus, MHA, Feliciano Yu, MD, MSPH, Phillip Asaro, MD, Stanley Paul Hmiel, MD, PhD

1St. Louis Children’s Hospital, St. Louis MO, and
2Washington University School of Medicine, St. Louis MO

Introduction

At St. Louis Children’s Hospital, orders for radiology exams are initiated from three disparate EHR systems (best-of-breed architecture). Each system has different exam nomenclature, causing inconsistencies and inaccurate orders. This led to delays in care due to clarification phone calls and inability to locate the appropriate orders. In 2013, through a hospital-wide multi-stakeholder process improvement effort (IS, clinicians, radiology staff), the disparate EHR dictionaries were harmonized with the existing Radiology protocols.

Intervention

After reviewing this problem with many trainees, who are the primary placers of orders in the institution, a solution was found in a standardized approach to radiology ordersets. To achieve this, a hospital-wide effort was launched to standardize the naming convention across multiple EHR systems as well as communicate with the clinicians ordering the exam. Synonyms were removed as they added confusion especially for trainees. Rather than an unending list of synonyms, the chief residents wanted the trainees to learn one standardized list of terms to use in the institution for radiology studies. Radiologist, orthopedists, radiology technologists, and pediatricians met on a regular basis to review terminology and agree upon naming standards. Through consistent naming of exams, trainees were forced to learn a specific set of terminology. To facilitate learning this set of terminology, ordersets were developed to assist order entry and facilitate education of these terms.

A radiology orderset was designed, initially for MRI orders and then CT and ultrasound orders. The order set was organized into a grid with rows and columns. Each row was a specific radiology study and there were 3 columns after each study for without contrast, with and without contrast, and with contrast studies. Studies were organized by clinical anatomy from head to toe. A sample is shown in figure 1.

Figure 1. Partial sample of MRI orderset organized anatomically and by contrast requirements.

Results

This intervention resulted in a 23.6% decrease in orders requiring corrections by radiology technologists. This rate of improvement was consistent for MRI, CT, and ultrasound orders. Currently approximately 30% of all CT, MRI and ultrasound orders are placed through these ordersets.

Conclusion

This approach to radiology orders was highly successful in both improving the accuracy of initial orders and teaching common terminology to the users based on usage rates. We believe this organizational strategy for radiology ordersets could be applied at other institutions as it was highly intuitive user interface that met the needs of the clinicians. We plan locally to expand our work to x-ray studies and other hospitals in our system.

References

Using National Database for Autism Research (NDAR) privacy-preserving record linkage protocol in the PEDSnet CDRN

Toan C. Ong, PhD¹, Michael G. Kahn, MD, PhD¹, Dan J. Hall²

¹Department of Pediatrics, School of Medicine, University of Colorado Denver, Colorado USA
²National Database for Autism Research, National Institute of Health, Bethesda, Maryland USA

Introduction: A national Pediatric Learning Health System (PEDSnet) is one of eleven Patient-Centered Outcomes Research Institute (PCORI) funded Clinical Data Research Networks (CDRNs), [1] In PEDSnet, a data-coordinating center (DCC) integrates clinical data from eight children’s hospitals nationwide into a centralized data warehouse and transforms the integrated data to the PCORNet common data model. PCORNet CDM v3 requires the inclusion of data sources other than the local electronic health records. So PEDSnet is integrating its current clinical dataset to external data sets by research subject. Because datasets are distributed among different institutions, a secure and efficient record linkage protocol is needed. A secure record linkage process must satisfy two security requirements: 1) No clear text personally identifiable information (PII) is sent to the DCC and 2) the hashing methods must be secure to protect against PII re-identification attacks.

The National Database for Autism Research (NDAR) project has developed a secure subject record linkage protocol based on centralized globally unique identifiers (GUIDs). [2] A GUID is a randomly generated unique string. The NDAR protocol comprises multiple steps, including locally encrypting PII using multiple one-way hash functions, transferring the hash codes to the NDAR server to assign common GUIDs to the same individual or to generate new GUIDs, and integrating distributed datasets at the data-coordinating center using the NDAR GUIDs. A brute force GUID generation attack is mitigated by stalling the user’s system response for 30 seconds following each 50 GUIDs returned and the GUID generation is continuously monitored. The current version of NDAR GUID tool generates SHA-512 hash with added salts, making collisions extremely rare, using a fixed set of linkage variables in multiple hash combinations allowing for common human factors entry errors to be identified. A future version of the NDAR linkage algorithm will be enhanced to incorporate more linkage variables.

Methods:

Discussion: The NDAR record linkage protocol can satisfy PEDSnet record linkage needs because this protocol 1) securely encrypts PII data locally using advanced hashing methods, 2) supports linking data from multiple partners and 3) gives the GUID back to the data partners. However, the current NDAR GUID generating tool has to be updated to separate the data encryption process from the data transferring process to transparently ensure that only hash codes are transferred to NDAR server. NDAR linkage algorithm must be modified to accept additional linkage variables, which will be used in case mandatory linkage variables are not available or incomplete.

References
1 Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc JAMIA Published Online First: 12 May 2014. doi:10.1136/amiajnl-2014-002743

Abstract: A National Pediatric Learning Health System (PEDSnet) integrates its clinical-based dataset with external data from various data partners. Because the datasets are distributed among different institutions, a privacy-preserving and efficient record linkage (PPRL) protocol is needed. A workflow to adopt a PPRL protocol developed by the National Database for Autism Research (NDAR) project is proposed to support these objectives.

Funding: PCORI CDRN-1306-01556
Best Practices and Design Experiences in Health Information Technology (HIT) Systems: A Primer

Gloria A. Opoku-Boateng¹, MS, Anthony F. Norcio¹, PhD
Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD, USA

Background: In healthcare where almost all Health Information Technology (HIT) systems are large and complex (based on tasks, users and environments), failure in design can result in both economical and human life losses.

Aims: There is the need to: re-evaluate the importance of planning in planning and system design, review recent successful and unsuccessful HIT system designs to find out how planning and design (alongside other practices) play a role in system success and positive user experience, and re-analyze design trends and come up with recommendations for better HIT system design.

Methods: This research is an exploratory study which is based on literature review of best practices, survey of subject matter experts and review of recent complex HIT systems for the above-stated research needs.

Results & Recommendations: This paper suggests six recommendations for incorporating best practices in system design of HIT software systems: 1. The need to fully understand the HIT domain; the healthcare, design and software development parts. 2. The need to understand the role and place of User-Centered Design and testing in HIT systems. 3. The need for a design team lead and/or expert in HIT (re-)design projects to help balance fuzzy logic versus structural IT logic. 4. The need to clearly distinguish system design vs user-experience for both development teams and users/clients. 5. The need for comprehensive analysis of any HIT system through planning. 6. The critical need for future work in the research and design process of HIT systems.

Conclusions: This paper is a call to arms to the HIT Systems community. A call to involve proper design processes and methodology in HIT systems. One cannot over-emphasize that fully planning and implementing the design of a HIT depends totally on a design team. Going through the system design cycle just like it is done for other complex IT systems does not cut it considering the risks and complexities involved in Healthcare IT systems. There is no doubt that a generally accepted convention for system design is to include both the feasibility and the requirements analysis phases but this should be an active phase where not only how viable a system is considered but how well do we know the current situation and proposed solution. Future research work is needed looking at bigger samples and also to test some of the theories and recommendations to see how they hold up.

Keywords: Health Information Technology (HIT), Software Development Life Cycle (SDLC) Feasibility Analysis, Requirements Analysis, User-Experience (UX), System Design.
Nurses’ Use of Electronic Health Records to Document Symptoms in Inpatient Settings: Preliminary Systematic Review Results

Mustafa Ozkaynak, PhD¹, Blaine Reeder, PhD¹, Lilian Hoffecker, PhD, MLS¹, Mary B. Makic, RN, PhD, CNS, FAAN¹, Karen Sousa, RN, PhD, FAAN¹
¹University of Colorado | Anschutz Medical Campus, Aurora, Colorado, USA

Introduction
Symptom management is one of the essential functions of nurses in inpatient settings. Electronic Health Records (EHRs) should support nurses in evaluating and monitoring symptoms. However, design and implementation of these technologies are critical factors related to the quality and quantity of this support. Moreover, suboptimal design and implementation of EHR technologies can hinder rather than facilitate the quality and timeliness of nurses’ documentation. Therefore, the purpose of this systematic review is to characterize the use of EHRs for documentation of symptom assessment and management in inpatient settings, and determine whether EHRs are more commonly used for specific symptoms. Understanding nurses’ use of EHRs for symptom management and documentation can give designers and researchers an opportunity to identify obstacles and facilitators to, as well as develop strategies to address, perceived electronic charting challenges.

Methods
The third author (a health sciences librarian) searched the Ovid Medline database (1946-2013) using MeSH and “free text” keywords that represented the concepts of electronic medical records, symptom documentation, and inpatient setting. For example, search terms for electronic medical records are represented by words, phrases, and acronyms like "chart," "medical records systems, computerized" and "EMR”. Space limitations precluded inclusion of all MeSH search terms. The last two authors provided nursing science and symptom expertise.

The search process yielded 274 studies. Studies that met inclusion criteria (a) were peer-reviewed; (b) explicitly mentioned EHRs; (c) their use by nurses; (e) to document patient symptom assessment and management; (e) in inpatient settings. Studies that were excluded described chart reviews but not the documentation process, physician-only documentation processes, nurses’ documentation processes but did not explicitly mention use of EHRs and/or were conducted in settings other than inpatient.

The first two authors conducted successive independent reviews and reconciliation of titles and abstracts, and full texts of downloaded articles. Reconciliation of title and abstract reviews resulted in 214 excluded studies, and full text reviews resulted in 50 excluded studies. Ten articles published from 2003-2012 were included. We conducted a preliminary analysis to identify symptoms, settings, challenges, and efforts to improve documentation.

Results
The most frequently documented symptoms were pain, breathlessness, fatigue, anxiety, nausea; and vomiting. Study settings described in included articles were post-surgery, oncology and palliative care settings. Reported challenges included: (a) the need for complex documentation of each symptom type; (b) lack of communication between staff; (c) fatigue while charting; and (d) inability of patients to fully communicate their symptoms (particularly pediatric population and those at end of life). Poor documentation results (e.g. missing essential details, lack of timeliness) were cited as a factor that led to patient safety and quality of care concerns. Interventions to systematize symptom management and documentation were cited as decreasing these concerns.

Conclusion
Full compliance with documentation of patient symptoms using EHRs in inpatient settings faces numerous barriers. EHR design should simplify documentation processes and decrease nurses’ data entry burden. Systematic approaches that account for individual, organization, task, and technology-related factors, and the interplay among them, could support nurses’ work and improve symptom management and documentation. The unique contribution of this review is that it narrows the gap in knowledge related to EHRs and nursing documentation. Full analysis of the completed systematic review will be presented at the AMIA Annual Symposium 2015.
Informatics Strategies to Address Cancer Worry of Urban Dominicans

Alsacia L. Pacsi, RN, DNS, CEN, CCRN, FNP¹ and Suzanne Bakken, RN, PhD, FAAN, FACMI¹,²
¹School of Nursing and Department of ²Biomedical Informatics, Columbia University, New York, NY

Abstract:
As part of the Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project, we conducted a survey that identified cancer as a worrisome health concern of urban Dominicans residing in Washington Heights/Inwood in New York City. Those who were married, younger, less depressed, and those who had received a cancer diagnosis previously were more likely to be worried about cancer. Survey data serve as a foundation to develop targeted informatics strategies to address cancer worry.

Introduction:
There is a paucity of studies centering on the correlates of cancer worry among Latinos from the Dominican Republic and the potential informatics strategies to address such worries. As part of the Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project, we conducted a survey that identified cancer as a worrisome health concern of urban Dominicans residing in Washington Heights/Inwood in New York City. The aim of this study was to generate potential informatics strategies based upon the correlates of cancer worry.

Methods:
The WICER sample comprised 3,079 Dominican respondents. Data were analyzed using descriptive and correlational statistics, and logistic regression with the dependent variable of cancer worry. Independent variables for the regression were: age, gender, marital status, education, socioeconomic status, previous diagnosis of cancer, anxiety, depression, sleep disturbance, and chronic burden. Potential informatics strategies were generated through brainstorming.

Results:
Most respondents were female and 63.9% identified cancer as a worry. Four variables significantly increased cancer worry: married marital status (OR = 1.19 [95% CI: 1.01, 1.40]), younger age (OR = .992 [95% CI: 0.987, 0.997]), less depression (OR = .96 [95% CI: 0.945, 0.982]), and cancer diagnosis (OR =2.269 [95% CI: 1.310, 3.931]).

Conclusion:
Study findings suggest that a high level of cancer worry exists in the community and that it is important to target messages about cancer based upon marital status, age, and whether or not someone has had a previous cancer diagnosis. This has several implications for informatics strategies. In terms of cancer worry, it may be useful to offer informatics-based tools in Spanish that will assist with calculation of risks for various cancers and culturally-relevant information related to level of risk. Given the Latino cultural value of familialism, it is not surprising that those who were married worried more. This suggests that informatics strategies should target cancer prevention for families, not just individuals. Personalism is also an important Latino cultural value. Thus, web-based video testimonials regarding cancer prevention that are targeted by age may be considered to be more relevant. The fact that those who had a previous cancer diagnosis were more worried suggests the need for tools that support cancer survivorship. Given that Latinos in general are most likely to access the Internet through their mobile phone than a desktop, it is important that cancer-related informatics strategies are suitable for mobile technology. Moreover, the relevance of SMS as an informational approach is also important given the predominance of basic rather than smartphones. The goal being to decrease cancer worry and empower the community to play more proactive role in making informed decisions and in turn promote their well-being.

Acknowledgement:
This research was supported by the Agency for Healthcare Research and Quality (R01HS019853, 1R01HS022961). Dr. Pacsi is supported by Reducing Health Disparities through Informatics (T32NR007969) from the National Institute of Nursing Research.
Privacy Preserving Sequential Pattern Mining Across Multiple Medical Sites

Doyel Pal  
Ph.D student  
Computer Science Department  
Oklahoma State University, Stillwater  
doyelp@cs.okstate.edu

Tingting Chen  
Assistant Professor  
Computer Science Department  
California State Polytechnic University, Pomona  
tingtingchen@csupomona.edu

Johnson Thomas  
Associate Professor  
Computer Science Department  
Oklahoma State University, Stillwater  
jpt@cs.okstate.edu

Introduction: Sequence Pattern Mining discovers frequent sequential patterns in large sequential databases. Sequence pattern mining methods are applicable in a large number of applications and domains such as, analysis of DNA sequences, patterns in patient path, etc. In these applications, survey of medical data from multiple parties is important which leads to more accurate sequential pattern analysis. When doing sequence pattern mining across multiple parties, under the regulations of HIPAA, patient’s privacy should not be revealed. Therefore we need to propose new privacy preserving schemes to perform sequence pattern mining across different parties. Our proposed scheme can be applicable in healthcare industries to identify the patterns in patient path, patient health, disease, medication etc. as it involves multiple parties and needs to be done in a privacy preserving manner.

Problem Statement: The goal is to find out the most frequent sequential patterns held by two or more parties without disclosing any sensitive data to each other. Given n parties, each party I has a database D_I. The SLPMiner algorithm [1] is used to find frequent sequential patterns at each site, e.g., P_1, P_2, ..., P_L. Some pattern P_l may appear at different sites. Denote P_l^A the appearance of P_l at site A, and P_l^B at site B. \sigma(P_l) is the number of times of P_l being selected as a frequent sequential pattern at different sites. If it is over a threshold, P_l is one of the most frequent sequential patterns across the participating sites. Privacy preserving sequential pattern mining reveals only the most frequent sequential patterns to other participating sites. Any other original data is kept secret from other sites.

Overview of the Methodology: The overview of our methodology is shown in Figure 1. The main steps are listed below.

1. Each site applies SLPMiner algorithm to find the local frequent sequential patterns.
2. Each site encrypts the result of Step 1 using public key.
3. All sites send the encrypted frequent patterns to a randomly selected site R.
4. Site R calculates Jaccard similarity coefficients [2] between any two encrypted patterns, and calculate \sigma(P_l) for each P_l as \sigma(P_l) = [(A,B)](P_l^A, P_l^B) = 1, where A, B \in \{1, n\}. 
5. If \sigma(P_l) \geq \tau, P_l is considered as most frequent sequential pattern. The plaintexts of E(P_l)s are eligible to be shared/published among all sites.

Evaluation Results: We evaluate our solution using different datasets generated by the IBM Quest research group [3]. We used two datasets with different numbers of sequences (2,000 rows of data to 10,000 rows of data). Each sequence of data consists of eight itemsets on average. Fig. 2 shows the efficiency performance of our solution for Data – Time to get frequent patterns between two sites.

Conclusion: We propose a new privacy preserving sequential pattern mining algorithm across multiple sites. It keeps each site’s data hidden from other sites except the mining results.

References
Organizing Drugs in RxNorm by Therapeutic Classes

Matvey B. Palchuk, MD, MS1,2, Michael Kamerick1
1TriNetX, Inc., Cambridge, MA; 2Harvard Medical School, Boston, MA

Abstract

The therapeutic classification of drugs is used ubiquitously. In the public domain, NDF-RT’s “Drug Products by VA Class” provides the ability to organize medications mapped to RxNorm. We used it to group RxNorm ingredients by therapeutic class. The result, while useful for analysis, is missing 13% of the valid RxNorm ingredients.

Introduction

The adoption of RxNorm for representing medications is gaining momentum. However, RxNorm does not address one of the key use cases for working with medications – classification. NDF-RT, a “sister” standard developed by the Veterans Administration (VA) is distributed alongside RxNorm and goes a long way toward addressing this need, but it falls short in certain areas.

Clinicians and researchers commonly think of medications in terms of therapeutic classes. We are very used to classes like “antibiotics” and “beta-blockers.” A prototypical therapeutic classification that is certainly familiar to every clinician is found in the Epocrates app for mobile devices (Epocrates, Inc., San Francisco, CA), where users see therapeutic classes in an intuitive hierarchy of several dozen top-level nodes (e.g., Antimicrobials, Cardiovascular) and limited depth (only three levels). Organizing medications in this fashion facilitates electronic prescribing, use of medication information for reporting and business intelligence, and research.

While many software applications with user-facing medication information find it necessary to represent drugs in hierarchical structures, there is a paucity of therapeutic classification schemes in the public domain. The “Drug Products by VA Class” from NDF-RT is the most prominent1, but it is uncertain how frequently NDF-RT maintainers update this classification and its mappings to RxNorm; this presents a challenge in relying on this classification system.

Methods

Our goal was to group RxNorm ingredients by therapeutic class using NDF-RT. We downloaded RxNorm (http://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html; accessed July 2014) and extracted the information into a relational database. NDF-RT therapeutic classification was derived from “Pharmaceutical Preparations/Drug Products by VA Class” parent node (NUI N0000010574). We traversed child-parent relationships to build our therapeutic classes hierarchy and “crossed over” into RxNorm at the level of clinical drugs (concept types SCD, SBD, GPCK and BPCK). Finally, clinical drugs were transformed into their corresponding ingredients using a “return the related concepts of specified term types” API call (http://rxnav.nlm.nih.gov/RxNormAPIREST.html).

To evaluate completeness of coverage, we obtained a list of all valid ingredients from RxNorm using a “return the RxNorm concepts for the specified term types” API call. We made certain that an ingredient had a corresponding clinical drug and excluded the ones that did not. We then compared the resulting ingredient list with the terminal nodes of the therapeutic class hierarchy we constructed.

Results

There are a total of 10,429 valid ingredients in RxNorm. The resulting NDF-RT therapeutic classification does not include 1,354 (13%) of them. This may be a reflection of the VA’s patient population – older and predominantly male.

Conclusion

Therapeutic classification in NDF-RT is very useful for organizing medications mapped to RxNorm, but it is incomplete. 13% of valid RxNorm ingredients – a significant number of medications – are not accounted for in this classification. Given how important therapeutic classification is for medication information, we would like to see it better addressed by standards in the public domain.

References

Perceptions of Health Information Technology Risks by Hospital Physicians

Sari Palojoki, M.Sc¹, Lasse Lehtonen, MD·PhD² and Kaija Saranto, PhD, RN, FACMI, FAAN¹University of Eastern Finland, Kuopio, Finland¹; University of Helsinki and Helsinki University Hospital District, Finland²

Introduction

Organizations can reduce the number and severity of electronic health record (EHR)-related safety events by anticipating the risk factors and an analysis of EHR-related safety concerns is a prerequisite for recognizing safety threats¹–².

We conducted a study to find out physicians’ perceptions of health information technology (HIT) risks. We identified all physicians employed in 23 different hospital facilities in Finland. Of these 2837 physicians, 650 were ineligible to participate in the survey because of absence from work or because they were not using EHR. Of the eligible respondents, 438 completed the survey, which yielded a response rate of 20%.

Materials, methods and results

A questionnaire based on Sittig and Singh’s theory and study findings¹–³ was constructed using a mixed-methods approach in several phases. A panel of experts in technology assessment, scientific background, clinical use of EHR and administration evaluated the face and content validity. Chronbach’s alpha was used for measuring internal consistency. The questionnaire consisted of eight set of questions, each with about five related questions about the design, development, implementation and the use of HIT. Statements were formulated with a five-point risk scale (from low [=1] to extreme risk [=5]). All responders were educated to use a standard risk matrix as a part of the hospital’s patient safety program.

The basic statistical analysis was performed by weighted average using the number of occurrences of each value as the weight. The highest perceived risk was the unavailability of the EHR (mean risk score 3.13). The second highest risk was in the failure to identify, find or use the most recent patient data. System-to-system interface errors were perceived as a high risk area e.g. due to the lack of direct interfaces between organizations (2.85). The highest risk scores were given by physicians working at emergency departments, operating rooms and intensive care units. The lowest risk was considered to be on the incorrect patient identification (2.39).

Figure 1. Mean risk scores of perceived HIT risks

Conclusion

HIT pose risks to patient safety. In our study, the physicians estimated that the HIT error types caused either a moderate or a high risks to patient safety. The study brings new data about perceived EHR risks: unavailability of the patients’ medical records, errors caused by miscommunication between applications and failure to find or use the most recent data were the most critical EHR safety concerns.

References

Measuring HIE in States and Factors Associated with States’ Success in HIE

Shriram Parashuram, PhD, MPH1, Seung Kim, PhD, MPH1,
Tianne Wu, MPH1, Prashila Dullabh, MD1
1NORC at the University of Chicago, Bethesda, MD

Research Objective
The State Health Information Exchange (HIE) Cooperative Agreement Program, enacted under the American Recovery and Reinvestment Act, expands the secure movement of electronic health information. As part of a multi-year evaluation of the program, we developed a composite measure of HIE progress for states in 2011, 2012, 2013. We studied the association between the composite state HIE measure in each program year, with program variables and pre-program contextual variables, to identify the key factors driving HIE in states.

Study Design
Development of a Composite State HIE Measure: We identified three priority domains of HIE: lab results exchange between hospitals and physicians, clinical care summaries exchange between physicians, and e-prescribing. From the American Hospital Association, National Ambulatory Medical Care, and SureScripts annual surveys in the three years, we identified provider-level measures that aligned with the three priority domains, and aggregated them to the state level for fifty-one states. After employing imputations to account for sampling variation in the three survey years, we standardized state-level measures and aggregated them domain averages for each state. The domain averages were appropriately weighted based on prior trends in adoption, complexity of exchange, and program goals, and aggregated into a composite HIE measure for each state.

Measuring factors associated with State HIE: Based on previous evaluation activities, we identified seven state program factors (e.g. laws to promote HIE under program, state-led HIE, availability of directed exchange services etc.) and eight state pre-program contextual factors (e.g. state size, population density, managed care penetration, level of hospital competition, pre-HITECH adoption of EHRs by hospitals and office based providers, etc.), conceptually associated with state HIE in the program years. We ran two sets of multivariate ordered logit models to study state contextual and program associated with states being in the highest HIE quartile: (i) multivariate models with contextual and program factors; and (ii) separate multivariate models with program factors and each endogenous program factor.

Principal Findings
States varied in their composite HIE measure across the three program years with changes attributable to state contextual factors in the early program years and program factors in the latter years. Smaller states, states with lesser competition among hospitals, and higher pre-HITECH office based-EHR adoption were likely to be in the highest HIE quartile. States that used query-based HIE with opt in consent models were likely to be in the lower HIE quartiles.

Conclusions
The results suggest that program factors became more important over the course of program implementation in their association with state HIE performance.

Implications for Policy, Delivery or Practice
Our objectives were to develop a state-level measure of HIE that was methodically robust, easy to compute and understand, accounted for program priorities, and was actionable, in that it allowed states to identify domains needing improvement. The composite measure also enabled us to study important contextual and program factors where states may want to direct their energies.
**Efficacy of Clinical Alerts to Decrease the Incidence of Contrast Induced Nephropathy**

Mark A. Parkulo, MD, Julia E. Crook, Ph.D, Launia J. White, B.S., Colleen S. Thomas, M.S. and March Rucci, M.H.A. Mayo Clinic, USA

**Introduction:** Contrast Induced Nephropathy (CIN) is a common and potentially preventable iatrogenic complication. We implemented Clinical Decision Support rules at Mayo Clinic in Florida between 2008 and 2010 to assist providers in identifying high risk patients and suggest alternative imaging or prophylactic measures to reduce the risk of CIN.

**Methods:** We compared the 12 month time periods before and after implementation of alerts with respect to the primary outcomes: (1) the proportion of patients with a creatinine test in the 30 days prior to imaging, and (2) the proportion of high risk patients receiving contrast. We secondarily examined the rate of CIN in patients for whom we had appropriate data. Fisher’s exact tests were used to compare the two time periods; graphical summaries were also used.

**Results:** Identification of high risk patients as evidenced by measuring of creatinine within 30 days of contrast use increased significantly for outpatients, both overall from 70% to 82% (P<0.001) and when focusing only on those who received contrast (see figure). However, there was no evidence of a reduction in the use of contrast in high risk patients. There was also no evidence of a drop in incidence of CIN, though this analysis was limited by (i) the fact that CIN incidence could be assessed only in those patients with appropriate pre- and post- creatinine measures, and (ii) the overall low CIN incidence.

**Conclusions:** The alert designed to increase the rate of creatinine measurement within 30 days of a potential contrast study was highly effective and sustainable as noted in the figure above, but the alerts designed to reduce the rate of contrast use were ineffective. This highlights the difference between actionable or nearly automatic alerts, such as ordering creatinine, as compared to alerts designed to inform or direct practice. The former alerts tend to be much more effective than the latter.
Building a custom lexicon for a large number of related concepts using templates

Olga V. Patterson, PhD\textsuperscript{1,2}, Matthew Freiberg, MD\textsuperscript{3}, Scott L. DuVall, PhD\textsuperscript{1,2}

\textsuperscript{1}VA Salt Lake City Health Care System; \textsuperscript{2}University of Utah, Salt Lake City, UT; \textsuperscript{3}Vanderbilt University, Nashville, TN

Abstract

Term recognition is the task of detecting the boundaries of a natural language expression in text that represents an underlying concept. Accurate term recognition relies on the availability of a comprehensive knowledge base that includes a representation of all text variations for a concept that can be potentially found in text. A data-driven lexicon generation results in a lexicon that more accurately matches the target corpus than an expert-driven approach.

This poster presents a method of developing a lexicon for a relatively large number of related concepts using patterns for a concept-value extraction of measurements in echocardiogram reports.

Background

Echocardiogram procedures result in a relatively large number (up to eighty) of heart function measures. In cases where these measures are not recorded in structured format, extraction from text of the report is needed. Concept-value pair extraction is a natural language processing (NLP) task that involves identifying occurrences of specific concepts in text and detecting the specific values mentioned in the document. In the case of extracting heart function measures, each measure is treated as a concept. The initial design challenge in creating such an NLP system is to identify the boundaries of terms for each concept. Abbreviations, spelling variations, word order variations, and misspellings create a large set of terms for each concept. An experienced clinician can create a list of some of these terms based on his/her personal knowledge because they are common. However, misspellings or less common variations require the list of possible terms to be determined by analyzing string occurrences in the text.

Methods

In the scope of a larger project aimed to extract heart function measures from clinical documents, a subset of the cohort was manually reviewed. It was noted that a large proportion of echocardiography reports followed one of several templates. While each of the templates used different words to express the same concepts, most templates contained similar patterns of concept and numeric value co-occurrence. These patterns are expressed as a linear sequence, in which a concept is followed by a short string of connective text and then by a numeric value with or without unit of measure, having one pattern per line of text. The text that appeared between terms and values was noted to be limited to a small set of expressions and to mostly contain punctuation or simple phrases. Since the number of the detected templates was quite large, manual annotation of all concepts in all identified templates was not feasible. However, templates prevalent in echocardiogram reports provided an opportunity for a semi-automatic acquisition of terms. We created an NLP pipeline that was able to extract potential terms from text. The system contains modules for: a) identifying phrases using regular expressions for finding numeric values with units and ranges of numeric values with units referred to as values [“0.4 mg”, “approximately 35%”]; b) identifying connective phrases using regular expressions that mark text typically found between terms and values [“is”, “=”, “!”, “found to be”]; c) using specialized annotation patterns that combine connective and value phrases using manually constructed patterns [“: 0.4 mg”, “ found to be approximately 35%”]; d) processing notes to identify strings in the concept slots as term candidates.

Results

A set of 445,487 documents was processed, and the term extraction resulted in a list of over 830,000 expressions. Most of these expressions were not related to any heart function concepts. However, the top 1000 terms of the most frequently encountered expressions were in fact representations of one of target concepts. The list of most frequent expressions was evaluated and a custom lexicon was compiled to link them to the concepts they represent.

Conclusion

Semi-structured documents allow for implementation of innovative approaches to building lexicons. The manual annotation burden can be eased leveraging templated sections of clinical documents through semi-automatic methods of term extraction.

Acknowledgements

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204 and the University of Utah.
Surveying Problem List Perceptions and Use in the Electronic Health Record

Darlene Peterson, MD1,3, Timothy Burdick, MD, MS2, Joshua Cohen3, Indra Neil Sarkar, PhD, MLIS3,5, Julie Lin, MD3,4, Dennis Plante, MD3,4, Elizabeth Chen, PhD3,5

1Maine Medical Center, Portland, ME; 2Oregon Health & Sciences University, Portland, OR; 3University of Vermont College of Medicine and 4University of Vermont Medical Center, Burlington, VT; 5Brown University, Providence, RI

Introduction
With the increased adoption of Electronic Health Record (EHR) systems, there has been renewed interest in the problem list and improving its use in the EHR. Several reports have described challenges for implementation and maintenance of problem lists, including the need for standardized coding, content, and use1,2. The goal of this survey-based study was to gain a better understanding of problem list perceptions and use at the University of Vermont Medical Center (UVMMC) for informing system improvements, user training, and policies.

Methods
A survey was designed and implemented using REDCap (Research Electronic Data Capture). Recruitment emails were sent to medical students and providers (physicians, nurse practitioners, physician assistants, fellows, and residents) at UVMMC in February 2013 and data collection was completed in March 2013. For closed-ended questions, basic statistics and graphs were generated; for open-ended questions, analysis involved developing and applying coding schemes for characterizing the contents of the free-text responses and identifying common themes.

Results
A total of 191 responses (~10% response rate) were received of which 160 (84%) were complete. Respondents represented a range of specialties with family medicine and primary care internal medicine as the most frequent; inpatient and outpatient settings; and, primarily attending physicians and residents/fellows. When asked to indicate where a particular condition should be documented in the EHR (problem list [PL], past medical history [PMH], past surgical history [PSH], family history [FH], social history [SH], or other), a variety of responses were provided for the 21 specified conditions. For example, 73% selected PL, 12% PMH, 73% SH, and 1% other for “ongoing tobacco use.” Several themes emerged from coding free-text responses to questions regarding problem list definitions, perceived barriers, and suggested improvements (κ>0.8 was achieved by two reviewers for a subset of responses). Overall, responses for definitions included the words “active” or “current” as well as referred to the problem list’s utility in healthcare decision-making. Some responses advocated for listing PSH, FH, or SH that have been deemed “important.” Overarching themes for perceived barriers included use of PL versus PMH, inconsistent use by providers leading to inaccuracy, and lack of ownership by providers. Overarching themes for suggested improvements included identifying one provider as “keeper of the list” (e.g., PCP) and building “hard stops” in the EHR to ensure that the list is updated by providers (e.g., at admission).

Discussion
The results from this survey are consistent with those reported in prior studies3,4. By evaluating both where in the EHR respondents would document specified conditions as well as their open-ended responses for problem list definitions and perceived barriers, varying perceptions and uses were observed. These findings provide further insights and guidance for standardizing problem list use, training, and policies.

References

Acknowledgments: This work was supported in part by National Library of Medicine grant R01LM011364.
Meta-Analysis of Ontology Applications in Healthcare

Kanitha Phalakornkule, MIS, MS, PhDc; Suhila Sawesi, PhDc; Josette Jones, RN, PhD
Indiana University, Indianapolis, IN

Abstract
This poster presents a meta-analysis of ontology applications within the last 15 years to evaluate the patterns of implementation processes and outcomes. The results showed that many studies needed to adjust the design of Ontology models in order to reach their long time goals. Upper ontology was absent in most of the applications even when their purposes was reuse and sharing. The role of upper ontology is to provide semantic interoperability of ontologies across applications (4). In addition, the majority of ontology implementations remain at strict research level and their long-term implementation plans were missing.

Background
Ontology is used to represent the conceptual structure of a specific domain by structuring its knowledge in hierarchal format. There is no right or wrong in ontological designing; however, the effectiveness of application is affected by the ontological knowledge model (1). Nowadays, the roles of Ontology in Information Science have been changing from an individual application like knowledge discovery, artificial intelligences (AI) and decision support systems (DSS) to a set of applications involving standardization, integration, reusability and interoperability connecting various unrelated applications to one another (2). The future of Ontology in healthcare tends to grow with the need of standardization and interoperability. Nevertheless, the current stage of Ontology in healthcare still requires more collaboration and understanding among applications in the domain (1) (3).

Method and result
To our knowledge, this is the first meta-analysis on ontology applications in the healthcare domain. In this study, our purpose is to analyze how Ontology has evolved in healthcare research and to discover what the outcome in Ontology implementation is. The PubMed was searched for studies published between 2000 and 2014 and potentially relevant studies were evaluated for eligibility. A total of 293 articles were met the predetermined inclusion criteria. During the data-analysis process, the patterns of ontology evolution in healthcare and the relations between their key attributes were evaluated.

In this meta-analysis, the purpose of integration is the most common (157 studies, 55.58%), while the purpose of reusability and prediction were the least common (19 studies, 6.48%). However, there were only 27 studies (9.22%) based on upper ontology models. Moreover, there were only 20% of the studies with the integration purpose were based on the upper ontology models.

The expectation is that this meta-analysis will lead us to better understand and to help us discover Ontological projects’ strength and weakness by presenting the relation between their proposes and their methodologies of the applications.

References
A Simulation Framework for Longitudinal Electronic Health Records Data

Matthew Phelan, M.S.¹, Benjamin A. Goldstein, PhD¹
¹Duke Clinical Research Institute, Durham, North Carolina

Abstract

Electronic Health Records (EHRs) are an increasingly utilized clinical data source. They are characterized by having multiple measurements on patients over time. Such longitudinal data present many analytic opportunities as well as challenges. These include: sampling analytic cohorts, modeling repeated measurements, and handling of missing and sporadic data. Having the ability to simulate realistic EHR data would allow researchers to study how different analytic approaches perform.

In this poster we present a simulation framework for EHR data. We model our simulation on EHR data derived from a large community hospital containing both inpatient and outpatient clinics. To motivate our example we consider laboratory measurements and biomarkers that are collected over time. We regard a patient’s presence in the EHR to be a function of several informative processes, namely an inpatient process and an outpatient process. We vary the rate of data collection depending on process state, mimicking the type of data one would observe in an EHR. Within this general framework, we consider such complicating scenarios as: data on multiple, correlated biomarkers, the presence of competing events, a cumulative effect of biomarkers on the probability of a given event, dynamic illness-to-recovery processes, and the presence of informative missing data. The final model is highly customizable allowing for varying degrees of prevalence of events, severities of illnesses and interactions among biomarkers.

Below, we illustrate how our simulation compares to real data. Figure 1a contains real glucose measurements on diabetics approaching a cardiac event, with a comparative control group. Figure 1b contains data from our simulated model illustrating how we are able to realistically capture the observed patterns.

Figure 1a

Figure 1b
Use of an Adaptive Agent-Based Model in Evaluating Patient Preferences in Healthcare

Andrew B. Phillips, PhD, RN\textsuperscript{1}, Margarita Sordo, MSc, PhD\textsuperscript{2,3,4}
\textsuperscript{1} MGH Institute of Health Professions, School of Nursing, Boston, MA; \textsuperscript{2} Clinical Informatics, Partners eCare, Partners Healthcare, Boston, MA; \textsuperscript{3}Department of Medicine, Brigham and Women’s Hospital, \textsuperscript{4}Harvard Medical School, Boston, MA

Introduction

The ability to effectuate change in healthcare increasingly relies on the skill of patients to make informed and intelligent healthcare choices. Whether these choices relate to the selection of providers, healthy behaviors, or insurance coverage, patients increasingly are at the center of healthcare reform efforts. Shared savings programs under the affordable care act reward improved patient outcomes through better patient management; meaningful use policies increase the information available to patients to manage their own care; wearable devices and the growth of mobile applications create additional sources of health data for patients. The goal is an “empowered” patient that efficiently navigates the healthcare system, making appropriate value-based decisions to reduce costs and improve individual and population health. New methods to model patient behavior are required for the design of technologies and policies that will successfully empower patients and transform care.

Methods

Complex systems science provides an opportunity to apply new tools to simulate and model patient behavior. A complex system describes a collection of agents that interact with each other to increase their individual health or fitness within a system. Each agent must make decisions based on individual preferences, values, and available alternatives within their environment. Due to the many interacting agents, complex systems do not behave linearly, but produce effects that are emergent and adaptive to changing circumstances\textsuperscript{1}. In the context of healthcare, patients can be considered one of the many agents interacting to maximize their health or fitness within a complex system. Three methods have emerged to simulate complex systems: system dynamics, network analysis, and agent-based modeling\textsuperscript{2}. System dynamics simulates system behavior through the operationalization of individual processes within a system to look at overall system behavior. Network analysis focuses on the relationships among agents within a complex system. Agent-based modeling simulates the complex system from the bottom up through the examination of individual agent behaviors based on discrete preferences, environmental factors, and interactions with other agents. Using agent-based methods we will develop a pilot computer-based adaptive model that identifies and operationalizes key health agents, agent properties, and interactions. Our model will potentially demonstrate the emergent behavior of patients in healthcare through the interaction of individual patients and other system agents.

Results

We hypothesize that an adaptive agent model will reflect the collective behavior and preferences of individual patients and their interactions with other agents in the system. By applying an adaptive agent model to simulate these patient behaviors patterns will emerge that potentially will improve our understanding of patient preferences, value perception, and decision-making.

Discussion

Healthcare continues to be modeled on provider-directed, evidence-based interventions that assume the patient is familiar and agrees on what is best for them – a provider centric top down model of care. The new roles envisioned for patients require bottom-up methods of evaluation to properly predict patient behavior. The development of an adaptive agent-based model allows for the consideration of the contextual, social, and other environmental impacts of the system on patients.

References

Data-driven identification of factors for appropriate selection of lab tests

Gouri Prakash, MISM, BE,
Independent Scholar, Mumbai, India

Abstract
The topic discussed in this paper is the systemic identification of lab tests that measure the same analyte from a data set, and the analysis and identification of heuristics that would guide the appropriate selection of lab tests.

Introduction
Enhanced decision support capabilities built into Computerized Physician Order Entry systems (CPOEs) at the point of entry of lab orders can ensure that clinician orders the right test at the right time for the right patient. It is no longer enough to give the clinician a laundry list of available lab tests to select from, in order to place the request for the lab test electronically vis-a-vis CPOEs. Building decision support, as an operational capability, involves identifying strategies that enable appropriate selection of lab tests at the point of entry for lab orders and LOINC can be leveraged to identify factors that enable the selection of appropriate test at the point of entry.

Methods
At the point of selection of lab tests, if two lab tests that measure the same analyte, are presented to the clinician, what factors drive appropriate selection of lab tests is the subject of this poster. LOINC has defined 5 parts for each lab test namely, component, scale, specimen, property, time and an optional sixth part, namely method. Component is the analyte of interest, or that biological, chemical or biochemical substance under measurement by the lab test. Scale indicates whether the lab test yields a qualitative or a quantitative result. Specimen is the source that contains the analyte, for example blood, urine, so on and so forth. Property refers to the property of the analyte under measurement, whether it is mass or volume concentration. The Top 300+ Commonly Ordered Tests hereby referred to as LOINC300+, which accounts for roughly 95% of all orders placed in the US, was downloaded from LOINC's official website and served to be the population of lab tests from which “matched pairs” were identified. Matched pairs are sets of two lab tests that measure the same analyte, however, exhibit variation in values for 0 or more for the other 3 LOINC attributes that characterize the test, i.e. Property, Specimen and Scale. All lab tests in the population had the value of ‘Pt’ (point in time) for Time and hence not being an attribute of discernment, it was excluded from the analysis. 41 panels present in LOINC300+ were excluded and for the 291 standalone tests, a JAVA program called LabTestAnalyzer was developed to extract ‘matched pairs’. Hence, a data-driven, systemic approach was devised to identify 2 lab tests that measure the same analyte from a standardized dataset. With the analyte (or Component), as the pivot, given the set of 3 attributes, {Property, Specimen, Scale} there were 2³ or 8 possible categories. Each matched pair was assigned to one of these 8 categories. Of the 24 matched pairs identified, 2 were excluded as 1 matched pair contained two tests which were identical and erroneously present twice on LOINC300+ and 1 matched pair had “XXX” as the value for Specimen. After the identification of the 22 matched pairs, the purpose and properties of each lab test were researched manually from the following sources of information on lab tests – (1) information provided in the Comment and Method_Typ LOINC supported fields (2) preamble available under the semantic category to which the lab test is assigned, in the Top 2000 Common Lab Test Results dataset (3) research papers in PubMed, indexed and retrievable by analyte name - to determine the differentiating factors that serve as a decision point for influencing test selection at the point of entry for lab orders.

Results and Discussion
22 matched pairs were identified which accounted for 44 of the 291 lab tests and each matched pair was assigned to one of 8 categories, for example, 7 pairs (31.8%) were assigned to Category 1, {Property, Specimen, Scale}, i.e. the values of all 3 attributes for both lab tests within the matched pairs assigned to this category were the same. Factors identified to be determinants for influencing selection between tests present within a matched pair were (1) purpose of test (2) sensitivity and specificity of test (3) analytical variation (4) processed versus unprocessed specimen (5) patient gender (6) different metric systems used in different countries (7) medication monitoring protocol as per clinical practice guidelines (CPG) (8) first line tests versus second line tests as per CPG. This poster highlights an application of data standardization different from interoperability - that of extraction of factors from standardized terminology to be used in the synthesis of lab test selection strategies at point of entry of lab orders.
Usability Testing of a Complex CDS Tool in the ED; Lessons Learned

Anne Press¹, Lauren McCullagh¹, MPH, Sundas Khan¹, MD, Salvatore Pardo⁴, MD, Andy Schachter¹, Thomas McGinn¹, MD, MPH

¹Department of Medicine, Hofstra North Shore-LIJ School of Medicine, Manhasset, NY, USA
²Department of Emergency Medicine, Hofstra North Shore-LIJ School of Medicine, Manhasset, NY, USA

Objective: Conduct usability testing to formulate an approach to successful implementation of the Wells criteria into the Electronic Health Record (EHR) in the Emergency Department (ED) of a large tertiary hospital.

Background: As the EHR becomes the preferred documentation tool across medical practices, healthcare organizations are pushing for clinical decision support (CDS) tools to be embedded into the EHR, to help bring evidence based medicine to the forefront of patient-physician interactions. However, often CDS tools are integrated into the EHR without an initial phase of usability testing, resulting in poor adoption rates. Usability testing is important because it evaluates an EHR CDS tool by testing it on actual users. An example of failed attempts to integrate CDS tools into the EHR, due to poor usability testing, is prior attempts to support appropriate decision-making regarding ordering CT (Computed Tomography) scans for the assessment of Pulmonary embolism (PE). This abstract outlines the usability phase of a study, which will test the impact of integration of a CDS tool for PE diagnosis into the EHR of a large urban Emergency Department, where workflow is often chaotic and high stake decisions are frequently made. We hypothesize that by conducting usability testing prior to the integration of the CDS tool into the EHR will result in increased physician adoption rates.

Methods: We conducted usability testing of a CDS tool that will be embedded into the ED EHR. The CDS tool consisted of the Wells rule for PE in the form of a calculator and was triggered off either CT orders or patients’ chief complaint and pertinent medical history. The study was conducted at a large tertiary hospital in Queens, New York. In addition to testing overall functionality, our specific usability question was whether a CDS tool with an earlier trigger (triage note) would be preferred to a later trigger at order entry. Seven residents were recruited and participated in two phases of usability testing. The number of participants recruited for this study was based on feedback from previous usability testing where a saturation of feedback was identified at approximately four participants. The usability testing employed a “think aloud” method, which called for participants to verbalize all thoughts as they interacted with the EHR and CDS tool, and “near-live” clinical simulation, where care providers interacted with standardized patients enacting a clinical scenario. Both phases were audiotaped, video-taped and had screen-capture software activated for onscreen recordings.

Results: Phase I: Data from the “think-aloud” phase of the study showed an overall positive outlook of the Wells CDS tool, in assessing a patient for a PE diagnosis. The subjects described the tool as “well-organized” and “better than clinical judgment”. Changes were made to improve placement of the tool into the EHR to make it optimal for decision-making, auto-populating boxes to minimize click fatigue and automation of the diagnosis into the documentation. Phase II: After incorporating the changes noted in Phase 1 participants noted improvements in the tool. There was less toggling between screens and the participants were able to complete patient visits more efficiently. Furthermore, it was clear from resident feedback that a triage note trigger assisted with decision making more so than the order entry trigger location.

Conclusion: This study successfully combined “think-aloud” protocol analysis with “near-live” clinical simulations in the usability evaluation of a CDS tool that will be implemented into a large emergency room
environment. Both methods proved useful in the assessment of the CDS tool and allowed us to refine the usability and workflow of the tool. The most effective point for triggering of the PE CDS tool was determined to be at the point of the triage nurse. In this manner we were able to prove the effectiveness of initial usability testing and its role in the implementation of CDS tools into the ED. Given the complex workflow of the ED, near-live testing and simulation environments are the most effective for this purpose.
Comparison of Knee Replacement Bundled Payment Pricing Variances Between Medicare, Medicaid, and Commercial Payers

Taylor R. Pressler, MS\(^1\), Thomas R, Gurgiolo, MA\(^1\), Scott Orme, PhD\(^1\), Robert Kershner, MPT MBA\(^2\), Jonathan D. Mathieu, PhD\(^2\),
\(^1\)Aver Informatics, Columbus, OH; \(^2\)Center for Improving Value in Health Care, Denver, CO

Abstract

Payment reform will include bundled payments. CIVHC is an all payer claims database within Colorado and presents a unique opportunity to compare costs and utilization patterns within an episode of care between payer groups. Episode analytics will provide an insight into providing clinically precise care that is financially efficacious.

Introduction

Bundled payments are defined as the reimbursement of health care providers (such as hospitals and physicians) "on the basis of expected costs for clinically-defined episodes of care"\(^1\). In 2013, Medicare started a pilot bundled payment program. Since then, several states, including Arkansas and Ohio, are starting bundled payment pilot programs for Medicaid populations. The Center for Improving Value in Health Care (CIVHC) is an all payer claims database for Colorado and offers a unique opportunity to examine claim and quality variances in care between different payer populations within the state. CIVHC is one of eleven such state claims repositories in the US. The focus of this paper will look at the differences between different payer groups (Medicare, Medicaid, and Commercial) and cost variations within a Knee Replacement episode.

Methods

The Knee Replacement episode was constructed using the Prometheus 3.6 definition. This definition provides codes and definitions for establishing which claims are part of a valid Knee Replacement episode, including which claims are considered Typical or Potentially Avoidable Complication (PAC). Episode results were compiled using the Aver Informatics platform and the results stratified by payer. In the case of dual eligibility, payer was determined by primary eligibility. Results to be included in the final poster presentation will include total allowed episode costs, average PAC cost, PAC proportion of total episode spend, and the ratio of PAC/Typical episode costs. Differences between groups will be compared using a Student t-test adjusted for multiple comparisons.

Results

Table 1 shows the results of the Knee Replacement episode. The Commercial payer grouping has the highest average episode costs, but the smallest proportion of PAC claims. This may indicate that providers have higher utilization of services during the episode. Additional analysis will be presented at the time of presentation to break out the differences in utilization patterns between payer groups.

Table 1. Results of the Knee Replacement episode comparing episode outcomes between payer groups.

<table>
<thead>
<tr>
<th>Payer</th>
<th>Total Episode Costs</th>
<th>Average Episode Costs</th>
<th>Average PAC Costs</th>
<th>Percent PAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicare</td>
<td>$7,172,911</td>
<td>$13,901</td>
<td>$3,359</td>
<td>24%</td>
</tr>
<tr>
<td>Medicaid</td>
<td>$7,763,544</td>
<td>$19,856</td>
<td>$7,564</td>
<td>38%</td>
</tr>
<tr>
<td>Commercial</td>
<td>$50,170,919</td>
<td>$28,013</td>
<td>$5,584</td>
<td>20%</td>
</tr>
</tbody>
</table>

Conclusion

Commercial payers have the least restrictions when it comes to the payment of fee for service claims. By utilizing the complete data within the CIVHC all payer claims database, differences in costs and utilization patterns can be examined between different payer groups for a large population. This will provide a unique opportunity for finding opportunities for optimizing payment reform by using a large claims data repository.

References


1636
Clustering Health Data to Discover EBP Interventions for Sepsis Prevention and Treatment for Health Disparities

Lisiane Pruinelli, MSN, RN1; Pranjul Yadav2; Andrew Hangsleben2; Kevin Schiroo2; Sanjoy Dey, PhD-C2; György J. Simon, PhD3; Maribet McCarty, PhD, RN4; Vipin Kumar, PhD2; Connie W. Delaney, PhD, RN, FAAN, FACMI1; Michael Steinbach, PhD2; Bonnie L. Westra, PhD, RN, FAAN, FACMI1,3

1School of Nursing, University of Minnesota; 2Department of Computer Science & Engineering, University of Minnesota; 3Institute for Health Informatics, University of Minnesota; 4Hennepin County Medical Center

Introduction
According to the Center for Disease Control and Prevention, sepsis or septicemia has doubled from 2000 through 2008, and hospitalizations have increased by 70% for this diagnosis. Patients with this condition are sicker, have longer hospital stays, and are more frequently discharged to other short-term hospital or long-term care institutions than patients with other conditions. Furthermore, there is a high rate of death or permanent organ damage, increasing health cost and burden. Despite the increasing number of new interventions and treatments, the challenge is early detection and prevention of sepsis and related complications. Use of evidence-based practice (EBP) guidelines, such as the Surviving Sepsis Campaign, could lead to earlier diagnosis, and consequently, earlier treatment of sepsis. However, these guidelines have not been widely incorporated in clinical practice. The overall goal of this study is to analyze patterns of EBP guidelines use and changes in health status over time, particularly for patients with health disparities.

Methods
After IRB approval, data from the EHR of a health system in the Midwest was transferred to a clinical data repository at the University of Minnesota established through a Clinical Translational Science Award. All episodes of care for patients hospitalized between 1/1/09 to 12/31/11 with a sepsis diagnosis were obtained and de-identified. Data included in the study were encounters, demographics, diagnosis, flowsheets, laboratories, medications, problems lists, and social determinants. Propensity score matching is used to compare a patient’s propensity in receiving guidelines interventions or not for different complications. Data mining techniques are used to create clusters of patient characteristics and to analyze the trajectories of these groups. Trajectories for EBP guidelines and groups conditions will be compared through survival curves.

Results/Discussion
The cohort contains 1,215 patients with ICD-9 codes for sepsis, SIRS, septic shock, laboratory associated sepsis, septic pulmonary embolism, and urosepsis. The risk for death for this sepsis cohort was 26.6%. However, when considering the risk for each of the ICD-9 codes, it was found that septic shock had the highest risk 46.9%, while septic pulmonary embolism had the smaller risk of death 11.3%. The majority of the patients are white (45.3%), followed by blacks (31.6%), and other ethnicities (23.1%). With respect the payer type, 13.44% were Medicaid, and 85.56% were no-Medicaid. Descriptive results comparing actual care to the EBP guidelines for the population as a whole and for those with health disparities, such as ethnicity and payer type, and other variables from the clustering analysis will be reported. Mining techniques applied to this cohort provide a better understanding of differences in predictive factors among septic populations that have different health outcomes. This study can contribute to better design and implementation of EBP guidelines to enable prevention, earlier detection and, better treatment of sepsis in specific disparate populations.

Acknowledgment
Support for this study is provided by the NSF grant IIS-1344135 and the National Center for Research Resources of the NIH 1UL1RR033183. Contents of this document are the sole responsibility of the authors and do not necessarily represent official views of the NSF, CTSI, or NIH.
Blood Flow Model for Improved Decision Support

Charles Puelz, M.A.¹, Beatrice Riviere, Ph.D.¹, Craig G. Rusin, Ph.D.²
¹Rice University, Houston, TX; ²Texas Children’s Hospital, Houston, TX

Description

This work is focused on improving clinical decision support and critical care provided to children impacted by Hypoplastic Left Heart Syndrome (HLHS). This congenital heart defect renders the left side of the heart useless; in turn, affected children undergo a sequence of three surgeries altering their cardiovascular plumbing to an abnormal state that functions with a single right ventricle pump.¹

We hope to improve critical care by building a mathematical model of this physiology, coupling the model with physiological data from Texas Children’s Hospital in Houston, Texas, and then interrogating the model with clinically relevant questions. In this way, clinicians and caregivers providing care to these patients at the bedside will be armed with enhanced knowledge about this complex physiology.

The mathematical model will be constructed with a network of one–dimensional elastic vessels representing the arterial and venous circulations. Zero–dimensional models for the organ beds and heart will be incorporated as boundary conditions into the vessel network. We will simulate these equations with a method similar to the one detailed by Sherwin et al.² Lastly, the data we will use for model calibration and development will come from the Catheter Lab at Texas Children’s.

References

Data Integration Opportunities and Challenges for the ADVANCE Clinical Data Research Network

Jon Puro, MPA¹, Vance Bauer, MA¹, Jennifer E. DeVoe, MD¹,², DPhil, Rachel Gold, PhD², MPH, Erika Cottrell, PhD, MPP¹, Andrew Brickman, PhD³, Christine Nelson, PhD, RN¹, Kenneth H. Mayer, MD⁵, Abigail Sears, MHA, MBA¹, Scott Fields, MD¹,², Pedro Rivera

¹OCHIN, Inc., Portland, OR., ²Kaiser Permanente Health Research, Portland, OR Inc., ³Health Choice Network, Portland, OR., ⁴Oregon Health and Sciences University, Portland, OR, ⁵The Fenway Institute; Boston, MA.

**Objective:** The Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Data Research Network (CDRN), funded by the Patient Centered Outcomes Research Institute, is “horizontally” integrating outpatient electronic health record (EHR) data for federally qualified health center (FQHC) patients, and “vertically” integrate hospital, health plan, and community data for these patients. The CDRN serves as a “community laboratory” for engaging vulnerable patients in Patient Centered Outcomes Research (PCOR).

**Design:** ADVANCE is engaging multiple stakeholders, focusing on three project pillars: the Data Pillar will develop and implement plans to normalize, validate, and expand a clinical data repository; the Cohort Pillar will develop and implement systems for engaging patients by collecting patient-reported data; the Regulatory Pillar will streamline existing policies for data security, privacy and confidentiality. ADVANCE’s work will also be supported by the Acuere Quality of Life (AcuereQOL™) data aggregation system to help FQHCs across the country achieve EHR Meaningful Use.

**Setting:** OCHIN, an Oregon-based health information network, Health Choice Network (HCN), a Florida-based health information network and Fenway Health, a Boston-based community health center.

**Participants:** ADVANCE CDRN includes FQHCs in 22 states that serve over 2.3 million patients, mostly women and children, low-income and either uninsured or federally insured.

**Results:** The increased data collection and integration supported by the ADVANCE Clinical Data Research Network will facilitate rich, impactful PCOR. Collecting and integrating multiple sources of data (ambulatory, hospital, claims, patient reported outcomes, and community data) raises many challenges. Some challenges are surmountable while others require workarounds or even changes of scope. Details of these challenges will be discussed.

**Conclusion:** Increasing numbers of FQHCs are embracing quality improvement and research as an integral part of care delivery. ADVANCE is committed to supporting these innovation efforts in FQHCs across the country, and ensuring that PCOR is aligned with the priorities of patients, clinics, and communities in our network.
To be discontinued: CPOE medications orders discontinued with reason being “error (erroneous entry)”

Arbor J.L. Quist¹, Thu-Trang Hickman, MPH¹, Alejandra Salazar, RPh, PharmD¹, Mary Amato, PharmD, MPH², Lynn A. Volk, MHS³, Adam Wright, PhD¹³, David W. Bates, MD¹³⁴, Gordon D. Schiff, MD¹³⁴

1. Brigham and Women’s Hospital, Division of General Internal Medicine, Boston, MA. 2. MCPHS University, Boston, MA. 3. Partners HealthCare, Information Systems, Wellesley, MA. 4. Harvard Medical School, Boston, MA

Background:
The Institute of Medicine estimated that over 1.5 million people are harmed by medication errors each year, making efforts to reduce medication errors important for improving patient safety¹. Implementation of computerized prescriber order entry (CPOE) systems has reduced medication error rates by creating legible prescriptions and incorporating clinical decision support, but these systems may also introduce new types of errors². The FDA funded this study to better understand how CPOE systems may cause medication errors.

Methods:
We examined a previously untapped database of more than 20,000 CPOE outpatient orders discontinued annually with the discontinuation reason “error (erroneous entry)” selected by the clinician. Retrospective review of over 180 instances of medications discontinued for error between 2011 and 2013 uncovered no documentation as to why prescribers made this selection, leading us to design a prospective study to capture first-hand information. We created a program that ran daily to email prescribers who had discontinued a medication clicking (err

Results:
During the four month prospective study period, we sent 542 emails to physicians. They responded to 58% of these query emails (n=312). We categorized the responses using a taxonomy classifying what happened to the patient and what went wrong in the CPOE system³. The top categories of “what happened in CPOE” included medication ordered for wrong patient (19.2%, n=60); wrong drug ordered (12.8%, n=40); and duplicate order placed (9.9%, n=31). Prescriber responses highlighted confusing aspects of the CPOE system and look-alike/sound-alike drugs, and also indicated the differing reasons for selecting “error (erroneous entry).”

Conclusions:
Data regarding erroneous medication entries represent an important source of “intelligence” about how CPOE systems are malfunctioning. Real-time queries to physicians placing discontinuation orders were remarkably more useful than chart review for rich details about reasons for errors in the original orders, with wrong patient and wrong drug errors being leading causes. Wrong patient errors may occur more in CPOE systems than non-CPOE settings as CPOE systems often allow prescribers to have multiple tabs open simultaneously for different patients. Prescribers also need education for better shared understanding of operational definitions for drug discontinuation check-box reasons. CPOE and clinical decision support developers may want to consider designing prompts for optional or required responses to provide details of the reasons when such a check-box category is selected. We found that most medications discontinued in error were discontinued within a minute of order, but future work should include full analysis of time between placing a medication order and its discontinuation to gain a better temporal understanding of when errors are identified and how CPOE interfaces play a role.

References:
Predictors of Medical Records Violation Punishments Filed with the Texas Medical Board (TMB) Before and After the HITECH Act
Mehdi Rais, PhD Student, Kavoos Eilami, MS Student, Craig W. Johnson, PhD
The University of Texas Health Science Center at Houston (UTHealth)
School of Biomedical Informatics, Houston, TX, USA

Introduction: Key Capabilities of EHR’s include health data and information management, results management, orders management, decision support, electronic communication and connectivity, patient support, administrative process and reporting1. Providers have adopted and are using EHRs in their practice settings at high rate since the seminal event of the passage HITECH Act in 2009. With wide-spread adoption of EHRs, changes have occurred in the way with which providers manage their medical records with these new technologies. Researchers hypothesized that such changes in the way with which medical records are maintained would have an impact on a subgroup of disciplinary actions around record keeping with state medical boards. To examine these differences in this virtually non-examined subgroup of disciplinary actions2, this research looked at physician violations due to inadequate medical records as filed with the Texas Medical Board between 2001 and 20143.

Methods: Cases targeted for the study were clinicians with an issued violation by the Texas Medical Board for Inadequate Medical Records occurring between January 2001 and October 2014. A sample of 1353 physician’s records containing characteristics for each individual that included age, gender, birthplace, medical degree, primary specialty, number of specialties, years in practice, and the type of disciplinary actions levied against the provider was obtained from the Texas Medical Board. From this data, it was determined that the two primary punishments issued by the Texas Medical Board were temporary period of probation or permanent license revocation.

Using SPSS 22, we performed hierarchical binary logistic regression and correlation analyses employing model building (1100), and cross-validation (253) samples. The dependent variable was the severity of punishment (Penalty Severity), with punishment being dichotomous (probation vs license revocation). The independent variables were age groups, gender, birthplace, degree type, before or after implementation of HITECH Act, primary specialty, and multi-specialty. Independent variable predictors were categorical in nature. The researchers reliably coded the occurrence/non-occurrence of license revocation versus non-license revocation for all issued violations.

Results: Using backward elimination, we identified age group, birthplace, and multi-specialty to be significant predictors for license revocation. Odds of a license violation recorded between January 2001 and October 2014 resulting in license revocation for inadequate medical records, controlling for other model predictors appear below:

- **Age Groups:** Clinicians in their 70’s had odds of license revocation 6 times those of clinicians in their 30’s; clinicians in their 50’s had odds of license revocation 3 times those of clinicians in their 30’s.
- **Birthplace:** Foreign born clinician had odds of license revocation 2/3 those of US born clinicians.
- **Multispecialty:** Multispecialty clinicians had odds of license revocation 1.6 times single-specialty clinicians.

Conclusions: The variables age, birthplace, and multispecialty have independent predictive value when evaluating license revocations for inadequate medical records. Findings were substantiated by significant (p < .05) odds ratios, Cox & Snell and Nagelkerke R² statistics, as well as significant (p < .001) R² statistics for models’ predicted probability of license revocation versus observed license revocation (Penalty Severity).

References
Automated Citation Retrieval System for Clinical Knowledge Management

Kalpana Raja, PhD, Andrew J. Sauer, MD, Melanie R. Klerer, Siddhartha R. Jonnalagadda, PhD

1Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL
2Bluhm Cardiovascular Institute of Northwestern Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

Abstract

Novel information retrieval methods to identify relevant citations on a clinical topic can overcome the knowledge gap between scientific publications and online clinical resources. We present a retrieval system that implements PICO framework, Vector Space Model and relevance based ranking module to support clinical knowledge management. The system achieved F-score of over 70% on preliminary analysis compared to baseline F-score of less than 5% and retrieved several new citations for new evidences.

Introduction and Background: The ever-increasing number of scientific publications has posed a challenge in maintaining up to date knowledge summaries in online clinical resources such as UpToDate. Our previous work to provide the best knowledge to clinicians prioritized the journals relevant to a clinical topic. Here we present a citation retrieval system for managing the clinical knowledge available in online clinical resources. Our approach retrieves both previously known citations and potentially new citations, which are not yet cited in online clinical resources, and all citations for new medical topics that are not present in online clinical resources.

Methods: An initial baseline, which is part of our existing system, for a given query is achieved using query expansion with NLM’s E-Utilities. A novel approach using PICO framework is introduced to filter citations with query related intervention or disease name in title and patient information in abstract. The relevance between MeSH (related to query) of citation and query is calculated using Vector Space Model. The citations are ranked based on Journal relevance, MeSH relevance, and VSM relevance.

Results: The input query is an UpToDate article and the output is a list of MEDLINE citations. The performance is evaluated using precision (P), recall (R) and F-score (F) on a gold standard (GS) consisting of citations from the article. The system achieved >70% F-score on a preliminary analysis with three UpToDate articles and retrieved a median of 6 potentially new citations as confirmed by a domain expert. The performance reported by the system is remarkably high when compared to baseline F-score of <5%. A very low precision of 1-2% due to large number of retrieved citations by the baseline decreased the F-score in spite of achieving 70-90% recall.

<table>
<thead>
<tr>
<th>Input query (UpToDate Article)</th>
<th>Citations in Gold standard</th>
<th>System performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Retrieved citations</td>
<td>New citations</td>
</tr>
<tr>
<td>Anticoagulation in older adults</td>
<td>74</td>
<td>71</td>
</tr>
<tr>
<td>Management of thromboembolic risk in patients with atrial fibrillation and chronic kidney disease</td>
<td>59</td>
<td>58</td>
</tr>
<tr>
<td>The use of antithrombotic therapy in patients with an acute or prior intracerebral hemorrhage</td>
<td>73</td>
<td>76</td>
</tr>
</tbody>
</table>

Discussion and Conclusion: Information retrieval using PICO framework as one of the modules achieved good performance. To our knowledge, PICO framework is used for the first time to filter retrieved citations. The analysis on UpToDate articles reveals that citations without abstract (PMID: 20962419) and MeSH (PMID: 24722803) are not identified by the proposed system. A performance evaluation with additional articles is needed for validating this work on automated maintenance of clinical knowledge in online clinical resources from scientific publications.

Acknowledgements: This work was funded by NLM grant R00LM011389. We are also thankful to UpToDate© for allowing us to use their data for research purposes.

References


1642
Closing the Loop with an Enhanced Referral Management System
Harley Ramelson, MD1, 2, Amanda V. Taube3, Pamela M. Neri, MS3
1Partners Healthcare Systems, Inc., Wellesley, MA, United States; 2Internal Medicine, Brigham and Women's Hospital, Boston, MA, United States; 3Clinical and Quality Analysis, Partners HealthCare, Wellesley, MA, United States

Introduction: Outpatient referrals made by primary care physicians (PCPs) involve multiple sequential steps requiring provider-to-provider and provider-to-patient communication. Analysis of malpractice claims has shown that the referral process in the ambulatory setting is prone to incomplete follow-up and communication break-down. An enhanced electronic referral management system (ERMS) was developed in an ambulatory EHR to meet a series of best practice steps for referral management. The objectives of this study were to assess PCP and staff satisfaction with the ERMS and to evaluate whether PCPs and their practice staff would find it easier to complete individual steps in the referral process and whether they would find it easier to identify when a problem has occurred in the process.

Methods: The ERMS was launched in two separate releases in 2014. The first release focused on functionality to create, transmit and track referrals as well as the ability to indicate that a referral was complete (“close the loop”) and to identify when a referral was overdue. The second release primarily focused on a series of reports allowing retrospective analysis of referral patterns. This study enrolled nine ambulatory primary care practices affiliated with an integrated delivery system in the Northeast. Approximately 100 practice staff members, including physicians, medical assistants, nurses, and administrative staff involved in the referral process were recruited across these practices to participate in surveys and interviews conducted pre-and post-implementation of both releases. Data reported here reflect results of the surveys conducted at baseline and following the first release. Ease of use and usefulness of specific functions were evaluated on a five-point Likert scale while satisfaction was evaluated on a seven-point scale.

Results: In the first 5 months following the release of the new ERMS module in the EHR, almost 3000 total referrals were generated. At baseline, 74% reported that it was easy or fairly easy to create a referral. This increased to 91% on the follow-up survey. At baseline, 33% reported that it was easy/fairly easy to schedule an appointment with a specialist. This increased to 64% on the follow-up survey. At baseline, the percent who reported that it was easy/fairly easy to identify when an appointment was not made with the specialist, when an appointment was missed or when an appointment was cancelled was 34%, 22% and 12% respectively. These increased on the follow-up survey to 54%, 45% and 36%, respectively. When asked about their overall satisfaction with the referral process, 36% of respondents reported they were satisfied or very satisfied at baseline compared to 92% at follow-up. The new functionality was also assessed in the follow-up survey to determine the usefulness of 15 major new functions in the module. Nine functions were found by 90% or more of respondents to be useful or somewhat useful. These include: 1. Ability to refer to a practice in addition to an individual specialist 2. Search functionality for a specialist 3. Favorite list of specialists and practices to whom PCPs commonly refer their patients. Supporting the intra-practice workflow of a staff person creating the electronic referral and sending to the PCP to authorize 5. Ability of the system to automatically fill in the appointment date based on interfaces with scheduling systems 6. Ability of the system to automatically link the specialist’s consult note to the referral based on the linkage with the notes module in the EHR 7. Ability of the system to identify when a referral is overdue at various stages in the referral life cycle 8. Ability to acknowledge the consult note in the referral module 9. Ability to document the closing of the referral. Conclusions: The results from the surveys after the new release of the ERMS show there is an increase in user satisfaction with the referral process and that most of the functionality was considered to be useful by a majority of the respondents. Users of the module found that previously difficult tasks such as the ability to identify when an appointment with the specialist was not made, missed or cancelled became easier with the new module. The implications for quality of care and patient safety are clear: when practice staff are able to track referrals in a more effective and efficient way and be able to intervene when appropriate, fewer patients will experience delayed specialty care potentially leading to improved patient satisfaction and patient safety.

References
Problem list problems: A look into data integrity

Gilbert H. Ramirez, M.D., Hyeon-eui Kim, R.N., Ph.D., M.P.H., Robert E. El-Kareh, M.D. M.P.H., M.S.

1University of California- San Diego, San Diego, CA

Abstract
Maintaining accurate problem lists for patients can be important for patient care, driving clinical decision support and identifying appropriate cohorts for research. However, problem lists are often inaccurate and out of date. We took the problem list data from 32,965 patients in our medical system and described issues with their data integrity.

Introduction
Accurate problem lists in electronic health records can support fast and efficient clinical care, clinical decision support and identification of research cohorts. We sought to look at the data integrity of these lists at our institution.

Methods
We created descriptive statistics for the types of providers who entered and revised problems, durations of problems before resolution and number of duplicates on the problem lists. As an example of an acute issue, we describe how long “urinary tract infection” remained on problem lists.

Results
Our preliminary results show 4941 (15%) patients with duplicate problem list items, which may be a considerable underestimation as we did not account for similar ICD codes in the initial analysis. We also found that only 5% of problem list items were resolved overall. For urinary tract infection, the average time to resolution in the problem list was 859 days 95%CI (753.16 to 964.84).

Preliminary data on those modifying problem list data (Figure 1)

![Figure 1. Types of providers modifying data in the problem list.](image)

Conclusion
Our preliminary data shows that attending physicians update problems most often, but these lists are inaccurate or out-of-date for many acute issues. Further study of practice style and data entry habits will help identify solutions.
Automating Maintenance of Care Team Relationships from Electronic Health Administrative Data to Decrease Variability of Care Coordination using the Health Information Exchange Infrastructure

Pallavi Ranade-Kharkar, PhD, MS1,2, Greg Gurr BS3, Darren Mann BS1,
Sidney N. Thornton, PhD1,2
1Intermountain Healthcare, Murray, UT; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 3Agilex Technologies Inc., Chantilly, VA.

Abstract
Primary care physicians refer one in three patients to specialty care. Yet, care coordination between primary care and specialty care lacks adequate information transfer and referral tracking. We propose an intervention that analyzes electronic health administrative data, extracts and curates relevant care team relationships and uses the health information exchange (HIE) infrastructure to automate care coordination for specialty care referrals enabling rule-based, standardized clinical data transmission.

Introduction
An integral aspect of care coordination is the “exchange of information among participants responsible for different aspects of care”1. For effective care coordination, the health system must understand the who, when, what and how of information exchange. Automation can reduce variability associated with manual processes2. Communication between primary care physicians and specialists is often inadequate, late and/or missing in about half of the referrals3. Insufficient time and logistical barriers were cited as hindrances for less than desired level of communication. Poor referral tracking and information transfer can lead to wasteful re-referrals, delays in diagnosis or treatment, diminished patient and physician satisfaction and poor quality of care. Intermountain Healthcare (IH) has built multiple connections for health information exchange (HIE) with state and national-level partners with interposed services that provide a platform for authoring logical processes for bi-directional data transmission.

Method
We created a system to process real-time health administrative data from multiple source systems, to link relevant patient visits, and to create a list of potential recipients responsible for different aspects of patient care within our trust framework. We curate this list based on patient’s current visit reason and create a distribution list of care team members. We have extended the existing HIE infrastructure to: 1) create triggers by intercepting the multi-source HL7 ADT transactions; 2) derive a care team of clinicians based on the incoming data and a rules repository that defines what relationships are pertinent; 3) build subscription services for clinicians to specify their preferences for types of documents of interest and frequency of exchange; and 4) send the requested document to the members of the curated care team using the HIE infrastructure at a pre-determined time and/or interval based on the visit context.

Results and Discussion
Building relevant care team relationships from source systems in addition to our EHR enables care coordination across organizational boundaries. Our approach to repurpose the HIE infrastructure to automate the communication between primary care and specialty care for referrals reduces the variability of information transfer by providing just-in-time and consistent solution that does not require manual intervention. The ability to subscribe/unsubscribe delivery preferences gives autonomy to clinicians to fit the type and frequency of communication to their needs.

Conclusion
Rule-based automation of data transmissions supporting care coordination eases the operational burden of busy clinicians and offers a predictable workflow. Creation and curation of relevant care team relationships from past and current administrative data provides a reusable resource that can enable additional service development.

References
Non-Physicians E-Prescribe More than Physicians in a Pediatric Emergency Department

Mordechai D. Raskas, MD, EdM1, Paul Mullan, MD, MPH1, Brian R. Jacobs, MD2,
James M. Chamberlain, MD1, Aric S. Campling, MS, RN-BC2, and Shilpa J. Patel, MD1
1Emergency Medicine, Children's National Medical Center, Washington, DC; 2Bear Institute,
Children's National Medical Center, Washington, DC

Introduction

Electronic prescribing (eRx) directly to pharmacies is associated with increased fill rates, improved clinical outcomes, decreased medication errors, and cost savings1,2. The Centers for Medicare and Medicaid Services includes eRx in the meaningful use criteria used to evaluate hospitals and providers. Studies suggest that non-physician providers (NPPs) provide equal or better care compared with physicians3. However, adoption rates of eRx by NPPs compared to physicians are not known.

Objective

To determine if non-physician providers e-prescribe at higher rates than physicians in a pediatric emergency department (ED).

Design/Methods

Retrospective analysis of discharge prescriptions from 2 pediatric EDs between 12/2013-10/2014 after implementation of a commercial eRX platform. Prescriptions without a designation of print or eRx were removed from the analyses (1.3%). Bivariable and multivariable regression analyses were used to determine associations between provider type and eRx status. Covariates included shift, day of week, calendar quarter, location (hospital ED vs. satellite ED), and medication type. Analyses included physicians (pediatric emergency medicine (PEM) attendings, general pediatricians, and fellows) and NPPs (nurse practitioners and physician assistants).

Results

There were 45,752 prescriptions from 128 ED providers: 33% eRx and 67% printed. NPPs and physicians had eRx rates of 41% and 31%, respectively (OR 1.5, 95% CI 1.46-1.61). The difference remained significant after adjusting for covariates (aOR 1.6, 95% CI 1.56-1.73). Daytime shift (7am-3pm) (aOR 1.4, 95% CI 1.31-1.47), Tuesday (aOR 1.2, 95% CI 1.11-1.29), and satellite ED (aOR 2.4, 95% CI 2.33-2.54) were associated with greater eRx rates. Oral antibiotics and NSAIDs accounted for 47% of the total prescriptions and had eRx rates of 37% and 28% (aOR 1.6, 95% CI 1.54-1.74). PEM attendings prescribed 13,917 of the total prescriptions; those with 5 years or more experience eRx at 41% and those with <5 years experience at 31% (aOR 1.7, 95% CI 1.56-1.84).

Conclusion

Non-physician providers e-prescribe more frequently than physicians. This difference may reflect more rapid adoption of new technology by non-physicians and is an area that deserves further investigation. In addition, if these increased eRx rates by non-physicians prove true in multiple settings, it may suggest utilizing non-physician groups as champions of eRx and targeting additional educational efforts at physicians. Daytime shift, Tuesday, satellite ED, and PEM attending with 5 years or more experience were all associated with higher rates of e-prescribing. We had expected more senior PEM providers to have lower eRx rates and were surprised to find the opposite. Future studies could explore this issue and qualitatively assess the factors and medication types associated with different eRx rates.

References

Social Media use for Drug Repurposing: Understanding the consumer perspective

Majid Rastegar-Mojarad, MS12, Hongfang Liu, PhD1, Priya Nambisan, PhD2
1Department of Health Sciences Research, Mayo Clinic, Rochester, MN; 2Department of Health Informatics and Administration, UW-Milwaukee, Milwaukee, WI

Introduction
Drug repurposing is the process of discovering new indications for existing drugs and is becoming an important component of drug development as success rates for novel drugs in clinical trials decrease and costs increase. In the period 2007-2009, drug repurposing led to the launching 30-40% of new drugs. Typically, a new indication for an available drug is identified by accident. However, new technologies and huge amount of available resources enable us to develop systematic approaches to identify and validate drug repurposing candidates with significantly lower cost. A variety of resources have been utilized to identify novel drug repurposing candidates such as literature, clinical notes, and genetic information. Another valuable resource in drug repurposing is patient experience and incidence of consumer drug repurposing. Social media is providing a platform for patients to share their experience with their illness and medications. One example of such drug repurposing by consumers through social media, is the use of Stilnox, an insomnia medication for brain injury. These kinds of consumer drug repurposing happens in many online forums and good examples of such forums are PatientsLikeMe and WebMD, both allow patients to freely share their experience with their treatment and medications. However, there haven’t been any studies to understand how and what kinds of drugs are being repurposed by patients in these online forums. In this study, we are focusing on WebMD to identify all events of drug repurposing.

Materials and Methods
First a list of commonly searched drugs in WebMD is retrieved. For each drug, we retrieve 1) known indications from DrugBank 2) side effects from a side effect resource called SIDER 3) patients’ reviews in WebMD which contain at least one disease name (except known indication and side effect). The list of disease names is retrieved from UMLS and finding disease name in the reviews is done by co-occurrence search. For the drugs with low number of retrieved reviews, we manually analyze them. For the rest, we implement a code to automatically detect patterns such as “I use drug for disease”. Each review has three rates a) Effectiveness b) Ease of use c) Satisfaction, and some information about the reviewer such as age, sex, and how long on treatment. This information uses to classify and prioritize the reviews.

Results
We started a list of 180 top searched drugs in WebMD. The average of number of reviews for each drug is 751. The maximum number of reviews is for Cymbalta with 4433 reviews and the minimum number is 41 reviews for Hydrocodone HD. Our initial analysis showed in 1800 reviews (10 for each of the drugs), 472 unique disease names (some of them are false positive because of disambiguation), 2973 mentions of disease names. The most frequent disease name is depression with 133 appearances in the 1800 reviews. Some of the other frequent disease names are migraine, itching, rash, arthritis, and dry mouth.

Implications
When patient use social media to share their experience with medication, most of time they talk about effectiveness of medications or side effect. Sometimes they share experience of using a medication for not FDA-approved indication. In this study, we try to identify these comments and detect the new indications. There are some limitations in this approach. First, this type of findings is not for serious and long-term diseases. In fact, patients can’t recognize effect of unrelated medication on a long-term disease. Second, as the reviews are informal comments and they contain misspelling and grammatical errors, developing a Natural Language Processing (NLP) system to analyze the text automatically is problematic. The first step in developing the NLP system is implementing a disease named entity recognition (NER). For obtaining preliminary results, we used a simple dictionary-based NER, in which false positive rate are high because of disambiguation. Some examples of disambiguation are: ‘still’, ‘down’, and ‘little’ which could also be disease names. Solving this issue in formal text with proper grammar and spelling is relatively easier than informal reviews in social media written by non-expert people. We plan to use machine learning NER. To assess the feasibility of using social media for drug repurposing, we will review the comments manually.
Web-based Patient-Centered Toolkit: Demographics of Enrollment

Sucheta S. Ravindran, MS, 1 Anuj Dalal, MD, 1,2 Conny Morrison, 1 Julie Fiskio, 1 John Hanna, 1 Diana Stade, 1 Kelly McNally, 1 Eli Mlaver, 1 Patricia C. Dykes RN, PhD 1,2 1 Brigham and Women’s Hospital, Boston, MA, 2 Harvard Medical School, Boston, MA

Introduction: In busy acute care settings, the utility of bedside patient portals for engaging patients in their care has great potential for enhancing patient education, facilitating communication with providers, and improving the patient experience. 1 With a diverse patient cohort in mind, we implemented a web-based patient-centered toolkit (PCTK), optimized for standard-sized tablet computers (e.g., iPads). The PCTK was designed to be easily usable by patients and caregivers, regardless of prior exposure to tablet functionality. 2 In performing this analysis, we aimed to analyze patient demographics by enrollment status (enrolled vs. declined to enroll), and to identify correlations between patient participation and key demographics.

Method: Patients admitted to medical intensive care and oncology units were approached for enrollment during the study intervention period. Demographic information, including age, race, gender, and education level, was retrieved from the electronic health record. We compared average age of enrollees versus non-enrollees using the t-test method. We measured correlations between demographic markers and patients’ willingness to participate using Spearman rho test. We calculated frequencies of various educational levels for enrollees and non-enrollees. These data were analyzed to identify differences between patients who participated and those who declined.

Results: Of 410 patients approached, 147 patients enrolled and 263 patients declined to participate. The average age of all patients was 58.5 ± [15.98] years; 240 (58.5%) were male; 308 (75.1%) were Caucasian; and 102 (24.9%) were non-Caucasian. The average age of patients who declined compared to that of enrollees was 60.5± [15.6] versus 54.9± [16.1] years [p = 0.0001]. There was a weak but statistically significant negative correlation between age and participation in use of the bedside toolkit (Spearman rho = -0.167; p = 0.001; alpha = 0.01). One-hundred and seventy-eight (43.4%) patients were graduates of college or postgraduate school, 82 (20%) had some college or vocational schooling, 114 (27.8%) were high school graduates or had a GED, and 17 (4.1%) had some high school education. Of the 147 enrollees, 90 (61.2%) were graduates of college or postgraduate school, 30 (20.4%) had some college or vocational schooling, 24 (16.3%) were high school graduates, and 3 (2%) had some high school education or less. Of 263 who declined to enroll, 90 (34.2%) were high school graduates/GED, 88 (33.5%) were graduates of college or postgraduate school, 52 (19.8%) had some college or vocational schooling, and 33 (12.5%) had some high school education or less. The most common reason why patients declined to enroll was a self-reported lack of need (110 patients; 41.8%). Other common reasons for declining were that patients did not speak English (42 patients; 16%) and that patients felt the toolkit involved “too much technology” (15 patients; 5.7%). Ninety-six patients (36.5%) declined for other reasons.

Discussion: In designing technology-driven bedside tools, accessibility and ease of use are important considerations, especially given diversity in patient demographics. In spite of tailoring our tool with a diverse patient population in mind, lack of previous experience or familiarity with bedside tools and personal willingness to participate are potential barriers to using technology as a means of engaging in one’s plan of care. Our analysis shows that the largest proportion of patients who both enrolled and declined enrollment had the highest education level. This is consistent with our finding that the majority of the patients approached was more educated. It is thus apparent that an analysis of a larger and more representative patient population might generate further insight into the goals of this particular study. Our results, however, do shed some light on important considerations for the use of informatics as a tool for improved patient engagement. Though we found a negative correlation between toolkit enrollment and age, it was not as strong as anticipated; in fact, this finding indicates greater success at enrolling older patients than initially expected and suggests more potential for varied enrollment. With regard to education, our findings suggest that, particularly with less educated patients, perhaps presenting the toolkit as a means of more actively engaging in one’s care plan may result in greater willingness and an improvement in patients’ perceptions of their need to utilize the toolkit in the future.

Acknowledgments: The BWH PROSPECT study is funded by the Libretto Consortium supported by the Gordon and Betty Moore Foundation

References:
Gaussian Processes for interpreting Multiple Prostate Specific Antigen
measurements for Prostate Cancer Prediction

Narges Razavian\textsuperscript{PhD,1}, Saul Blecker\textsuperscript{MD,2}, David Sontag\textsuperscript{PhD,1}

\textsuperscript{1}Computer Science Department and \textsuperscript{2}Medical School, New York University, New York City

Background and Introduction:
Prostate cancer is the second leading cancer cause of death among US men. Prostate Specific Antigen (PSA) is currently the primary approach used for prostate cancer screening, but the recommended cutoff (\( \geq 4.0 \text{ng/mL} \)) is limited by low sensitivity\textsuperscript{1}. In this work, we present a machine learning approach based on Gaussian processes\textsuperscript{3} that is able to better use repeated PSA measurements in detecting prostate cancer. Gaussian Processes are semi-parametric methods, which use a distance measure to construct a kernel matrix and approximate a function from possibly noisy and irregularly measured observations. They enable inference of the function value at new data points, and provide confidence of the inferred values.

Methods:
We use an initial cohort of 4.1 million beneficiaries, enrolled in a major insurance company between 2005 to 2013. We included male beneficiaries age\( \geq 50 \) who had at least two PSA readings. The primary outcome was prostate cancer, defined by a prostate cancer diagnosis, AND either having at least one PSA measurement\( > 10 \text{ng/mL} \), or having gone through cancer treatment (bilateral orchiectomy, chemotherapy, cryotherapy, external beam radiation therapy, prostatectomy, radioactive seed implants, brachytherapy, or hormone therapy). A male patient was defined not to have cancer if no evidence of any prostate cancer (diagnosis or treatment) or PSA\( > 10 \) were ever observed in their records. We excluded patients who were not in our positive or negative class. We used as the index date (i.e., the “prediction time”) the first date at which a patient was observed to have a PSA of above \( 4.0 \text{ng/mL} \), or before the time of the cancer diagnosis (the earliest of these two dates). We used PSA values and velocity feature using slope of the \( \log(\text{PSA}) \) over the past years\textsuperscript{3}. We then use two methods of linear regression (in log space), and Gaussian processes(with linear and cubic kernel functions)\textsuperscript{2}, to transform PSA observations to a fully observed time series, and classify the time series using logistic regression and support vector machines. We train the models using randomly selected 60% of the data. We use 20% of the data to tune parameters of the model, and use the remaining 20% as the validation set.

Results and Discussion:
Of the patients that matched inclusion criteria, 1018 developed prostate cancer. Table 1 shows the quality of the cancer detection for different methods. Our results confirm previous findings that multiple observations of the PSA values improve the prediction sensitivity\textsuperscript{4}, while retaining high specificity. However, we conclude that interpreting the multiple observations is important. Velocity features\textsuperscript{3} help with the sensitivity, however they suffer from low specificity. By using Gaussian process (both linear and cubic) we can improve the downstream cancer classification accuracy. Two major classifiers (SVM and logistic regression) show similar quality for classification, with logistic regression model slightly better, but our major gain was achieved from better modeling of patient’s all previous PSA observations.

<table>
<thead>
<tr>
<th>Model</th>
<th>SPC</th>
<th>SNS</th>
<th>PPV</th>
<th>NPV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Had PSA ( = 4.0 )</td>
<td>0.87</td>
<td>0.64</td>
<td>0.85</td>
<td>0.68</td>
<td>0.83*</td>
</tr>
<tr>
<td>Had PSA ( &lt; 3.5 )</td>
<td>0.97</td>
<td>0.97</td>
<td>0.92</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>Logistic regression based on all previous measured PSA (without imputation)</td>
<td>0.73</td>
<td>0.83</td>
<td>0.77</td>
<td>0.80</td>
<td>0.85*</td>
</tr>
<tr>
<td>Velocity curve (linear in log space) as imputation + logistic regression on the inferred time series</td>
<td>0.84</td>
<td>0.80</td>
<td>0.82</td>
<td>0.83</td>
<td>0.90*</td>
</tr>
<tr>
<td>PSA imputation using linear Gaussian process + logistic regression on the inferred time series</td>
<td>0.85</td>
<td>0.80</td>
<td>0.82</td>
<td>0.83</td>
<td>0.89*</td>
</tr>
<tr>
<td>Linear Gaussian process + logistic regression on the inferred time series</td>
<td>0.87</td>
<td>0.81</td>
<td>0.87</td>
<td>0.80</td>
<td>0.88*</td>
</tr>
<tr>
<td>Linear Gaussian process + confidence intervals using linear Gaussian process + logistic regression on the inferred time series</td>
<td>0.84</td>
<td>0.79</td>
<td>0.83</td>
<td>0.81</td>
<td>0.89*</td>
</tr>
<tr>
<td>Linear Gaussian process + SVM on the inferred time series</td>
<td>0.85</td>
<td>0.80</td>
<td>0.82</td>
<td>0.83</td>
<td>0.90*</td>
</tr>
</tbody>
</table>

SNS= Sensitivity, SPC = Specificity, PPV=Positive Predictive Value, NPV=Negative Predictive

* All values have 95% confidence interval of size \( \pm 0.003 \)

References:
Identification of Variables that Predict Visit Times for Analyzing Ophthalmology Clinic Workflows

Sarah Read-Brown\textsuperscript{1}, Michelle R. Hribar, PhD\textsuperscript{2}, Grant Aaker, MD\textsuperscript{1}, Leah Reznick, MD\textsuperscript{1}, Thomas Yackel, MD, MPH, MS\textsuperscript{2}, Michael F. Chiang, MD, MA\textsuperscript{1,2}

\textsuperscript{1}Ophthalmology and \textsuperscript{2}Medical Informatics & Clinical Epidemiology; OHSU, Portland, OR

Introduction
Patient flows in clinical ophthalmology are complex and multifaceted which leads to the accumulation of patient wait times, affecting patient satisfaction\textsuperscript{1} and future care\textsuperscript{2}. Being able to predict variability and patient exam times has potential to greatly improve scheduling strategies\textsuperscript{3}. We propose using observational data, electronic health record (EHR) data, and provider’s perceptions of patients to identify variables that predict visit lengths.

Methods
This study took place at OHSU, in a single ophthalmology provider’s clinic (LR). Two authors (SRB, MRH) observed a clinic for 8 days, using time motion methods and electronic forms to record the times that patients spent during each part of their exam. After each exam, the provider was asked to categorize patients as complex vs. non-complex in regard to their medical diagnosis and encounter interactions with the provider. We then combined observed times and perceptions with timestamps and patient characteristics recorded in the EHR (Epic; Verona, WI). We calculated the length of the total interaction time with the provider and staff (total exam time) as well as the total length of the office visit (total visit time). The total visit time includes the total exam time plus wait times.

Results
Table 1 shows the total exam and visit times based on provider perception, age category, and dilation. The factor that most strongly predicted total exam time was the provider’s perception of a complex vs. non-complex encounter. The factor that most strongly predicted total visit time was dilation.

Table 1. Influence of various factors on mean total exam time and total visit time

<table>
<thead>
<tr>
<th>Provider perception</th>
<th>Total exam time (minutes)</th>
<th>Total visit time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
<td>Mean</td>
</tr>
<tr>
<td>Complex diagnosis</td>
<td>16 (24%)</td>
<td>30</td>
</tr>
<tr>
<td>Complex encounter</td>
<td>13 (20%)</td>
<td>37</td>
</tr>
<tr>
<td>Complex diagnosis and encounter</td>
<td>6 (9%)</td>
<td>36</td>
</tr>
<tr>
<td>Non-complex</td>
<td>43 (65%)</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infant (0-12 months)</td>
<td>3 (5%)</td>
<td>33</td>
</tr>
<tr>
<td>Toddler (1-4 years)</td>
<td>28 (42%)</td>
<td>23</td>
</tr>
<tr>
<td>School Age Young (5-10 years)</td>
<td>22 (33%)</td>
<td>26</td>
</tr>
<tr>
<td>School Age Old (11-17 years)</td>
<td>7 (11%)</td>
<td>29</td>
</tr>
<tr>
<td>Adult ( &gt;18 years)</td>
<td>6 (9%)</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dilation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>30 (45%)</td>
<td>26</td>
</tr>
<tr>
<td>No</td>
<td>36 (55%)</td>
<td>27</td>
</tr>
<tr>
<td>Total Visits</td>
<td>66 (100%)</td>
<td>27</td>
</tr>
</tbody>
</table>

Conclusion
Secondary use of EHR timestamps and observed clinical data can help understand workflow by predicting exam time and wait time, and may be used to increase office efficiency in scheduling. In addition to traditional patient demographics, provider perception may be a useful indicator of the total exam and total visit times.

References
Novel Template Identification from VA Text Integration Utility Notes

Andrew M. Redd, PhD1,2, Guy Divita, MS2,3, Adi V. Gundlapalli, MD, PhD1,2, Le-Thuy Tran, PhD2, Mathew Samore, MD1,2

1Division of Epidemiology, Department of Internal Medicine, University of Utah; 2IDEAS Center, VA Health Care System, Salt Lake City, Utah; 3Department of Biomedical Informatics, University of Utah;

Abstract
Templates in text notes pose challenges for information extraction. We are developing a system that identifies novel templates in plain text notes included in Veterans Health Administration (VA) Text Integration Utility (TIU) notes. The system is comprised of two parts. The first clusters documents according to groups that are likely to have been derived from the same template. The second attempts to identify elements of the template from these documents.

Introduction
For the purpose of this project we define template to be a set of textual elements where additional information is inserted by a care provider. A common example of this kind of template would be a form for recording a patient’s vital signs, heart rate, blood pressure, etc. Templates pose a particular challenge for information extraction (1). For example, templates can cause an excessive number of false positives (2). These challenges were the inspiration for developing a tool that identified templates without the assumption of knowing the form or content of the templates.

Previous Work
The most relevant work that relates to this are algorithms related to plagiarism detection including detecting copy and paste sections. Also a relevant area of research is document section identification which overlaps in the goal of trying to tag information to text.

Methods
The system we are developing separates the process into two parts. First we identify groups of documents that are similar, with the premise that these documents are likely to contain the same template. To accomplish this, documents are tokenized on whitespace, case sensitive with punctuation attached to the word if there is no whitespace between. To handle the large number of tokens possible, tokens are reduced using cyclical redundancy check as a many-to-one dimension reduction map, with 512 possible bins words could be grouped into. Hashed tokens are counted and a hierarchical clustering algorithm used to create groups. The logic behind this is that documents created from the same template will show similar token counts.

The second process takes groups of similar documents and extracts the relevant templated text. Under the assumption that templates are text with other text inserted, if all the documents in a group do indeed contain the same template, the common elements of all documents are likely to be the template. We implemented a longest common subsequence (LCS) to extract the elements of documents that were common to all documents and in the same order in all documents.

Results and Next Steps
We applied the methods to a 750 document corpus, a random selection of notes from many different note titles, and will present results. We are scaling the methods to run on a larger corpus with approximately one million documents. This will result in a new catalog of templates with statistics regarding use across documents and type of use, in addition to being able to be used as a tool for identifying documents that use the template. The long term goal is to complement ongoing efforts in improving information extraction techniques for VA big data.

References
Preserving Semantic Content of Narrative Clinical Information in the OMOP Common Data Model Format

Grant Reed, PhD and Carol Friedman, PhD
Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

To date, little research has been performed on how to preserve the semantic content of relevant clinical information that is contained in narrative clinical notes, while representing the information in the format of a common data model, such as the Observational Health Outcomes Partnership common data model (OMOP CDM).

Introduction

In order to detect adverse event signals, effective pharmacovigilance research requires a vast amount of clinical patient data. A common data format facilitates the analysis of large data sets from disparate data sources. For example, Observational Health Data Sciences and Informatics (OHDSI) researchers utilize clinical information originating from claims data and electronic health record data, which is mostly structured, that has been transformed into OMOP CDM format. However, common data models are typically constructed to accommodate fairly simple structured clinical data, not fine grained structured data obtained from narrative clinical information, which is semantically rich and includes many different types of modifying information. As a result, there is potential for loss of semantic content upon transformation into the CDM format.

Methods

We converted narrative clinical notes to structured content using MedLEE, which is a natural language text processor that captures comprehensive clinical information that includes qualifiers such as certainty, severity, temporal information and family history. We then established how to represent the rich structured form in the OMOP CDM format. One aim was to compare the data content of the structured data obtained from natural language processing to the type of data content that is accommodated by the OMOP CDM. A second aim, which is future work, is to tailor the extraction-transformation-load (ETL) in a manner that minimizes loss of semantic content.

Of particular interest is how to preserve the broad range of qualifiers found in narrative clinical text in the OMOP CDM format. For example, certain qualifier types of information can be stored in the OMOP CDM Observation table, as foreign keys to standard concept identifiers corresponding to qualifiers. However, what must be determined is whether or not the fields can accommodate the complexity and comprehensiveness of qualifier information contained in the rich structural form of MedLEE output. Also of interest will be to demonstrate that the fine grained data is useful in pharmacovigilance research.

Conclusion

This research should preserve semantic content in narrative clinical data upon transformation into the OMOP CDM format, and should be useful for increasing the amount and richness of clinical information available to OHDSI researchers. This research may also lead to recommendations for enhancing the CDM format to accommodate more complex data useful in pharmacovigilance research.

References


Adapting a Mobility Monitoring Protocol for Sensor Studies with Functionally Vulnerable Older Adults

Blaine Reeder, PhD¹, Angela Richard, PhD, RN¹, Jason Falvey, PT, DPT¹, Roger Paxton, PhD¹, Jennifer Stevens-Lapsley, PT, PhD¹
¹University of Colorado | Anschutz Medical Campus, Aurora, CO

Introduction
Mobility is a strong predictor of hospital readmission in older adults after acute hospitalization. However, little is known about the mobility levels of this population after discharge home. In-home sensor technologies have great potential to monitor mobility levels in support of independence for older adults. However, some older adults live with multiple morbidities that limit their capacity to engage in research. This challenge has resulted in few large-scale sensor studies that enroll older adults. Therefore, the purpose of this presentation is to describe how we adapted a mobility monitoring protocol previously employed with healthy community-dwelling older adults to reduce participant burden for functionally vulnerable older adults that enroll in sensor studies.

Methods
Adaptation of the mobility monitoring protocol to a short form was informed by experiences from a feasibility study with functionally vulnerable older adults who received a progressive multi-component (PMC) physical therapy intervention after discharge home from an acute hospitalization. The original mobility protocol totals 74 items and comprises physical, psychosocial, and cognitive parameters intended to correlate self-reported and sensor-based measures of mobility. We used the PMC protocol as criteria for instrument substitution to reduce the total number of participant-completed items in these parameters.

Results
Instrument replacements reduced the total number of items from 74 to 43, a total reduction of 31 items. Table 1 shows specific changes made for physical and psychosocial parameters. No changes were made for cognitive parameter measurement. Monthly paper fall calendars were replaced with a single weekly Y/N question about falls or near-falls using a tablet computer. Two semi-structured technology interviews (60-90 minutes), at study midpoint and end, were replaced with a 15-question technology acceptance survey and a short interview (30 minutes), to be administered at study end.

Table 1. Specific changes to mobility monitoring protocol and participant impact

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Original</th>
<th>Short Form</th>
<th>Participant impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical</td>
<td>15-item SPFM²</td>
<td>12 items: AM-PAC³ “6-Clicks” basic mobility &amp; AM-PAC³ “6-Clicks” daily activity assessments</td>
<td>15 fewer items, clinician assessment vs. self-report</td>
</tr>
<tr>
<td>Psychosocial</td>
<td>15-item GDS³, 19-item MOS-SS³</td>
<td>2-item PHQ-2⁴, 4-item PROMIS⁵ Informational Support – Short Form 4a</td>
<td>13 fewer items</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15 fewer items</td>
</tr>
<tr>
<td>Fall Calendar</td>
<td></td>
<td>Single Y/N question about falls/near-falls</td>
<td>Single question vs. form to complete</td>
</tr>
<tr>
<td>2 semi-structured</td>
<td></td>
<td>15 question technology acceptability survey &amp; 1 short interview</td>
<td>Up to 45 fewer minutes to complete</td>
</tr>
<tr>
<td>interviews</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

²Self-reported Physical Function Measures; ³Activity Measure for Post-Acute Care; ⁴Geriatric Depression Scale-Short Form; ⁵Patient Health Questionnaire; ⁶Medical Outcomes Study Social Support Scale; ⁷Patient Reported Outcomes Measurement Information System

Conclusion
Our next steps are to pilot the short form mobility monitoring protocol in a feasibility and acceptability study that tests the use of in-home sensors in support of a progressive multi-component physical therapy intervention for functionally vulnerable older adults. The authors would like to thank Drs. Hilaire J. Thompson and George Demiris from the University of Washington as the originators of the original mobility monitoring protocol.

References
Identifying Patterns Indicative of Copying/Pasting Behavior in Patient Generated Online Content

Tera L. Reynolds, MPH,1 V.G.Vinod Vydiswaran, PhD,2,4 Yuting Wu,3 Qiaozhu Mei, PhD,1,4 David A. Hanauer, MD, MS,5,1 Kai Zheng, PhD6,1

1School of Information; 2Department of Learning Health Sciences; 3College of Literature, Science, and the Arts; 4Department of Electrical Engineering and Computer Science; 5Department of Pediatrics; 6School of Public Health Department of Health Management and Policy. University of Michigan, Ann Arbor, MI

Background
Providing patients with ready access to, and ownership of, their clinical information (e.g., laboratory results, discharge summaries) through technologies such as patient portals has been a much anticipated paradigm shift in healthcare. However, existing research provides evidence of several barriers to the adoption and use of patient portals, with one of the primary barriers being low numeracy and poor health literacy among patients.1 Online health forums provide an opportunity for researchers to gain insights into the questions that patients ask about clinical information on peer-to-peer communication channels, which is indicative of the struggles that they may have in understanding the information. The types of questions posted in online health forums vary greatly; from requesting emotional support to posting detailed clinical information that appears to be copied and pasted or transcribed (referred to below as copying/pasting behavior) directly from a patient’s medical record (e.g., laboratory results). In order to further study the challenges that may arise when patients are presented with clinical information, we considered the latter type of question and, as an initial goal, focused on identifying instances of copying/pasting behavior related to laboratory test results.

Methods
We downloaded all questions and threaded comments from a popular online health forum, MedHelp.org, in February 2013. The full dataset contains 1.4 million questions and 4.6 million answers/comments generated by over a million unique users on various health-related topics. We utilized the live MedHelp.org website in the development and testing of various keyword searches (e.g., “Lab result help”) by viewing 90 posts of the results of each search for: (1) the number of instances of copying/pasting behavior, and (2) the “Related Discussions” section for additional words/phrases to test. We then translated the keyword searches with the highest proportion of relevant posts to complex queries in the Indri Query language.2 The queries were run against an inverted index built over all the MedHelp.org question posts. We manually reviewed the top 20% (600) of the retrieved results to identify patterns indicative of patient copying/pasting behavior. During the review of the last 100 results, no additional general patterns were added, suggesting saturation had been reached.

Results
Among the MedHelp.org question posts with evidence of copying/pasting behavior (n=184), seven general patterns were identified: (1) 93% specific test name and numeric result (e.g., TSH 0.11), (2) 47% lab normal/reference range, (3) 38% “results below”, (4) 27% units (e.g., 1000.0 IU/L), (5) 23% specific test name and categorical result (e.g., HCV RNA: Positive), (6) 16% tabular, and (7) 4% quotation marks (e.g., “Herpes simplex type 2 isolated”).

Conclusion and Future work
We identified several patterns indicative of patient copying/pasting behavior in a popular online health forum. We intend to continue this work by: (1) comprehensively extracting all instances of copying/pasting behavior, (2) searching for themes in the extracted posts in an effort to better understand the underlying questions in the posts, and (3) based on this understanding, suggesting ways to improve patient portals to help patients better understand their data. In addition, future research may include developing methods to identify instances of copying/pasting behavior from other sources of clinical information (e.g., discharge summaries and clinical notes).

References
An Analytics Approach for Adverse Drug Event Discovery

Mohammadreza Rezaie1, Kenji Yoshigoe, Ph.D.2, Umit Topaloglu Ph.D.1
1Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR; 2Department of CS, University of Arkansas at Little Rock, Little Rock, AR

Abstract
Healthcare organizations generate massive quantities of data which are often convolved with valuable hidden insights. In order to discover new adverse drug events (ADEs) the data mining techniques are employed extensively, but the computational complexity of the traditional data analytics platforms makes the need to migrate to a scalable analytics platform. The goal in experiments was predicting the possibility of certain interactions given a prescribed set of drugs. In this work, we used Spark platform and MLlib library as an alternative scalable solution that we found them scalable, easy to work with and with reasonable modeling accuracy.

Introduction
Discovering the drug adverse effects in both trial and post marketing phase is important for patient safety and efficacy of drugs. Data mining and predictive analytics proposes a promising solution to deal with this task. From the predictive analytics perspective, the goal is to predict the possibility of different symptoms given a prescribed set of drugs. However, due to the massive scale of the dataset, the traditional analytics tools and methods are no longer applicable or very time consuming. Magnitude of data (e.g. the FDA Adverse Event Reporting System [FAERS] dataset) to be analyzed urges us to exploit big data technologies and platforms, such as NoSQL.

The FAERS dataset contains patients’ demographic and drugs usage history information along with the symptoms associated with their prescription. We propose a prediction model, which attempts to find correlations between drugs and drugs to symptoms. The model is then used to predict the chance of exposure to different interaction given a prescription. It also generates alerts in case that the probability is higher than preset threshold value.

Our approach
We utilized Apache Spark to process the data and MLlib to perform the predictive modeling task. Spark is a parallel computing framework that is well suited for processing the large scale distributed data on large clusters. MLlib is a machine-learning library developed on top of Spark and contains many widely used algorithms for predictive modeling and data mining. We prepared a dataset from the most seen interactions and drugs in FAERS dataset (the top 10 interactions and top 30 drugs, based on their popularity) and imported it into Spark with a readable format for MLlib.

The result of many experiments we conducted on the dataset, proved to be promising. We have run Naive Bayes, Decision Tree and Random Forest Classifiers on the FAERS dataset and the Random Forest performed better in accuracy with 75.99%. The measure of accuracy here is the mean of correctness of interactions predictions compared against the model trained over 60% of data, and tested on the remaining 40% in a cross-validation type of analysis.

Conclusion
Predicting possible interactions between drugs over the FAERS dataset is crucial for the post-marketing surveillance program. The large size of dataset can result in high computational complexity that might not be tolerable for traditional data analytics tools. That fact strengthens the argument for the use of Spark and MLlib as a reasonable choice for predictive modeling. Throughout our experiments, we found Spark and MLlib to be scalable, easy to work with and well documented.

References
Visualizing High Dimensional Clinical and Tumor Genotyping Data

Matthew J. Rioth, MD1,2, Jeremy L. Warner, MD1,2
1Vanderbilt Department of Biomedical Informatics, Nashville, TN; 2Division of Hematology/Oncology Vanderbilt University Medical Center, Nashville, TN

Introduction

The increasing use of next-generation sequencing in the treatment of cancer produces multiplying amounts of genotyping data1. However, the representation of population-level genotyping data within a clinical context for operational, quality-improvement, and research purposes is difficult given their high-dimensionality and poly-hierarchical representation2. We describe novel interactive visualizations of these data that make them more intelligible, intuitive, and useful.

System description

We utilized data from targeted exome tumor sequencing results transmitted from a 3rd-party laboratory, in a standardized extensible markup language (XML) format. These XML documents were parsed for clinical as well as genomic data and aggregated into a relational database to enable population-level depiction for secondary uses. The data contained in this database were depicted with interactive visualizations using off-the-shelf data analytics products as well as custom scripted interactive web-based documents. The utility of these visualizations are becoming apparent to clinicians, administrators, and researchers and we are continuously improving and refining them based on feedback and evolving needs.

Figure 1. Example interactive bubble-plot visualization of population level genetics across tumor types.

References


Automated Prediction of Human Mobility Patterns in International Humanitarian Response

Nicholas Robison¹,², Anne M. Turner, MD, MPH, MLIS¹,²
¹Division of Biomedical and Health Informatics, University of Washington, Seattle, Washington
²Department of Health Services, University of Washington, Seattle, Washington

Introduction: Over the past decade international response to humanitarian emergencies has increased in its operational complexity involving the collaboration of multiple organizations, governments, and private individuals to provide supplies, infrastructure, and aid workers to address the emerging crisis. However, the amount of information regarding population dynamics, available resources, and immediate needs, made available to these responders is severely limited and has not kept up with changing needs. One particulate challenge is in understanding the underlying dynamics and distribution of affected populations. After a disaster, populations go through a period of tremendous fluctuation and realignment as individuals move away from damaged areas. This dynamic migration creates challenges for providing necessary resources. Aid workers often find themselves relying on out of date population counts and maps, which are rendered obsolete in the face of mass population movements.

Description: With the rise mobile communications and powerful data processing systems initial efforts have been undertaken to evaluate the use of cellular communication patterns to plot and predict population movements in near real-time (¹–³), these studies have, through simulation, found an achievable predictive accuracy of around 88% (²); however, efforts to translate the theoretical findings into a working system suitable for deployment into a disaster situation are lacking. To address this need, we present Cascade, a unified analytics system designed to integrate call data records (CDRs), determine the geographic distribution of individuals within the affected areas, predict their future movement patterns, and present the findings in a user-friendly way that can facilitate decision making on aid distribution and allocation. The process begins with inputting CDRs, provided by the local telecom operators, into a dynamic Bayesian network (⁴), which is then used to make real-time predictions of incoming events streamed from the telecom network. This system allows for both robust machine learning training, and rapid analysis of emerging information and movement patterns.

Conclusion: By combining historical data and new events, and mapping them with an adaptive kernel density estimation (⁵), we can provide synthesized information suitable for use by emergency responders to better understand population dynamics and provide improved resource allocation.

Acknowledgements: This work was supported in part by the NIH National Library of Medicine (NLM) Training Grant Nr. 2T15LM007442 at the University of Washington.

References

The CDS Collaborative: Goals, Deliverables, and Future Directions

Salvador Rodriguez, PhD,1 Emory Fry, MD,2 Tadesse Sefer, MS,2 Phillip Warner, MS,1 Claude Nanjo, MAAS, MPH,2 Jerry Goodnough,2 David E Shields,1 Steven Elliott,2 Esteban Aliverti,2 Kensaku Kawamoto, MD, PhD, MHS1

1Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
2Cognitive Medical Systems, San Diego, CA

Introduction

Standards-based clinical decision support (CDS) holds significant potential for enabling CDS at scale across institutions. To enable such standards-based CDS, several organizations have been focusing on the development and dissemination of open-source, standards-based CDS tools and resources, including the OpenCDS initiative (www.opencds.org), coordinated by the University of Utah, and the Socratic Grid initiative (www.socraticgrid.org), coordinated by Cognitive Medical Systems. The tools and infrastructure established by these groups are in use in a number of healthcare systems and electronic health record (EHR) systems. For example, Cognitive Medical Systems is currently leading the development of CDS capabilities in the next-generation EHR platform of the Veterans Health Administration using OpenCDS as a core component. To accelerate progress in this area, the CDS Collaborative was established to serve as a common home for organizations committed to open-source, standards-based CDS. Here, we describe the goals, deliverables, and future directions of the Collaborative.

Goals and Deliverables

The goal of the CDS Collaborative is to accelerate improvements in health and health care through the leveraging of open-source, standards-based CDS tools and resources. The deliverables of the Collaborative include the following:

Defined High-Value Use Cases. Given that the end goal for the Collaborative is to improve health and health care, the Collaborative has defined an initial set of high-value use cases needed by its current implementation partners. Further definition of use cases is ongoing based on the identified needs of members of the Collaborative.

Common, Standards-Based, Service-Oriented Architecture. Collaborative members are establishing a common software architecture based on the loose coupling of functionally independent Web services compliant with standard service interfaces, including the Health Level 7 (HL7) Decision Support Service, HL7 Common Terminology Services, HL7 Unified Communication Service, and HL7 Event Publication-Subscription Service standards.

Encapsulation of Software Components in Cloud-Ready Containers. Individual services and other components of the overall architecture are being encapsulated in Docker (www.docker.com) containers to enable cloud-ready deployment with minimal need for configuration. Services and components that have been encapsulated as of June 2015 include the OpenCDS Decision Support Service, several knowledge authoring environments (e.g., those based on the JBoss Guvnor and the JBoss KIE frameworks), and terminology management solutions leveraging the Unified Medical Language System and the Apelon Distributed Terminology System. These cloud-ready containers can be deployed on local servers as well if desired.

Sand-Box Environment with EHR Integration and Synthetic Data. In order to provide a productive and hands-on environment for the development and validation of CDS tools and resources, the CDS Collaborative is establishing sand-box environments complete with sample EHR systems (e.g., VA VistA and CPRS), CDS tooling, and sample CDS knowledge content. Synthetic data generation mechanisms are also being used to populate representative patient records. These sand-box environments are being deployed on virtual servers in a manner that can be replicated easily onto other servers.

Future Directions

Moving forward, the CDS Collaborative will be focused first and foremost on improving the scope, depth, and value of the tools and resources available through the collaborative. The Collaborative is also reaching out to interested collaborators who share in the Collaborative’s vision for improving care through open-source, standards-based CDS.

Conclusion

The CDS Collaborative provides a community, tools, and resources for the scalable deployment of open-source CDS solutions to improve health and health care.
A Case Study on the Effectiveness of an In-house Physician Rating Tool in Outpatient Clinics

Liu Ruijue1, Jiang Junhui, BEng1, Nadee Goonawardene, BSc1, Sharon S L. Tan, PhD1
1Department of Information Systems, National University of Singapore, Singapore

Introduction: To improve service quality and patient experience, healthcare institutions are placing increasing emphasis on providing patient-centered care, which ensures that patient value, preference and needs drive clinical decisions. Recent years have seen a surge in third-party healthcare rating websites as well as an emergence of in-house rating systems that aim to provide evaluations and feedback on clinics and physicians. While the effectiveness of myriad quality and outcome measurements has been studied, research on evaluating the effectiveness of rating tools in healthcare setting remains scant. The aim of this study is to evaluate the effectiveness of an in-house developed physician rating tool. In particular, we examine how patients’ perceptions and concerns about the rating tool affect their responses and how physicians’ perceptions and concerns affect their acceptance and use of the rating tool.

Methods: A multi-method study was conducted at a private medical center providing out-patient specialist services. This medical center uses an in-house developed clinical rating tool for customer experience monitoring and physician rating. The rating tool was designed based on a survey of various physicians rating tools and brainstorming sessions with the research team. After each visit, patients are invited to provide their feedback on an iPad which involves giving an overall recommendation score, and rating several dimensions including doctors’ listening, explaining, consultation time, as well as staff behavior, clinical environment, obtaining appointment, waiting time, and price performance. Qualitative analysis was based on interview data of patients and physicians which explored how they perceived and used the tool. Interviews were conducted in five clinics: 1) LASIK, 2) Eye, 3) Wellness, 4) Orthopedic, and 5) Obstetrics and Gynecology (OBG). From each clinic, 5 to 10 patients who have used the tool and 1 to 2 physicians being rated were randomly invited for a face-to-face interview. Each interview, which was audio-taped with the permission of the participant, lasted around 10-15 minutes and each participant was given a $10 grocery voucher. The interviews were then transcribed, coded and interpreted by researchers. Quantitative analyses involved conducting linear regression on archival rating data from the rating system. Results from both quantitative and qualitative analyses were triangulated to inform our findings.

Preliminary Results & Discussion: Our preliminary findings provide some interesting insights. First, based on quantitative analyses of the rating data, patients’ recommendation score for physicians is highly correlated with their physician’s explaining ($β=0.260$, $p<0.001$), listening ($β=0.162$, $p<0.001$), time spent ($β=0.114$, $p<0.001$), waiting time ($β=0.080$, $p<0.01$) and price performance ($β=0.078$, $p<0.01$). Based on qualitative analyses of interview data, patients were found to value different healthcare service dimensions. For example patients in OBG emphasized more on the interactions with physicians whereas patients in LASIK emphasized more on price performance and waiting time. Such differences were found to be rooted in two underlying factors: the outcome uncertainty of the treatment (i.e. experience-oriented – when medical outcome is more certain versus outcome-oriented – when medical outcome is less certain) and the time-horizon of the clinical interaction (i.e. short-term vs. long-term).

Second, majority of patients felt the rating tool was useful and were willing to participate. However, some concerns pertaining to three types of question sensitivities – threat of disclosure, question intrusiveness, and social desirability existed. For example, in terms of the “threat of disclosure”, patients who chose to be anonymous were found to give a lower average recommendation score to their physicians than patients who left their email addresses ($β=0.111$, $p<0.001$). Interview results also corroborated with this finding where some patients shared that they might artificially inflate their ratings if their identities were being disclosed for fear of affecting the relationship with their physicians.

Third, physicians were found to differ in their acceptance of the rating tool depending on their attributions of the ratings they received. Based on the interview data, physicians who attributed low ratings to unstable, internal or relational causes (e.g. lack of understanding or physician effort) were found to be receptive towards being rated by patients and as they felt that ratings could provide useful feedback for improvement. Physicians who attributed low ratings towards stable and external causes (e.g. patient incompetency rate or tool deficiency) expressed resistance toward being rated as they felt that low ratings were due to factors beyond their control.

Our findings provide useful implications on how to design an effective physician rating tool and increase patients’ participating and physicians’ acceptance.

Reference

Use of Diagnosis Related Groups to Predict All-Cause Pediatric Hospital Readmission Within 30 days

Victor M. Ruiz, MS¹, Amie J. Draper, BS¹, Ye Ye, MS¹, Andrew Urbach, MD², Fuchiang (Rich) Tsui, PhD¹

¹Real-time Outbreak and Disease Surveillance Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
²Children’s Hospital of Pittsburgh of UPMC, Pittsburgh, PA

Abstract
This study assessed the contribution of diagnosis related groups (DRGs) in predicting all-cause pediatric 30-day readmissions. We used machine-learning techniques to identify predictive DRGs, and built a Bayesian classifier from a training dataset of 64,075 inpatient visits between 2007 and 2011. The model had an area under the ROC (AUROC) curve of 0.76 (95% CI 0.75-0.77). We believe that DRGs can be used to target post-discharge interventions aimed at reducing readmission rates.

Introduction
Unplanned 30-day hospital readmissions have been increasingly used as an indication of poor quality of care. Since 2012, the Centers for Medicare and Medicaid Services (CMS), under the Affordable Care Act, started penalizing hospitals with higher-than-expected rates of 30-day readmissions. These penalties caused hospitals to lose up to 1% of Medicare inpatient revenue in 2013.

Identifying patients at high risk of 30-day hospital readmissions is key for targeting costly care-transition or post-discharge interventions, aimed at lowering readmission rates. Unfortunately, predictive models have shown poor ability to identify 30-day all-cause readmissions.¹ In this study, we aim to assess the contribution of All Patient Refined (APR) and the CMS DRGs in predicting all-cause 30-day readmissions.

Methods
Data collection: We retrieved CMS-DRGs and APR-DRGs from inpatient visits to the Children’s Hospital of Pittsburgh of UPMC (CHP) between 2007 and 2013, excluding deaths during admission and visits missing DRG codes. Eligible visits comprised 13,645 readmission cases and 77,716 non-readmission controls.

Datasets: We used visits from 2007 to 2011 (9,547 cases and 54,528 controls) to train a naïve Bayes (NB) model, and visits from 2012 to 2013 to test its performance. A readmission-reduction program began at CHP in 2013; thus, we divided the testing data into two sets comprising visits from 2012 (2,119 readmissions and 11,676 controls), and 2013 (1,979 readmissions and 11,512 controls).

Modeling: We conducted a two-phase feature selection. First, we ranked features in the training dataset using the information gain score, which measures the reduction of uncertainty after observing an attribute. We discarded features with no discriminative ability. Second, we added features incrementally to a NB classifier in order of rank. At each step, we measured the 10-fold cross-validation AUROC after adding a feature to the current best model, and discarded those that did not increase the AUROC.

Results
The training dataset had 283 and 749 distinct APR-DRGs and CMS-DRGs, respectively. The best model included 14 APR-DRGs (along with their risk-of-mortality and severity-of-illness scores), and 7 CMS-DRGs. The model had an AUROC of 0.76 (95% CI 0.75-0.77) on both testing sets, and a Brier skill score (calibration metric) of 0.12.

Conclusion
We predicted pediatric 30-day hospital readmissions with high accuracy from DRGs. Although DRGs are available after discharge, they can be used to target post-discharge interventions aimed at reducing readmissions. Future work will include an assessment of the predictive power of data available before discharge, followed by a pilot at CHP to validate our prediction strategy in a production environment.

References
Representation of Genetic Variants in Genomic Sequencing Reports

Evelyn Rustia, MD1,2, Chunhua Weng, PhD2, Carol Friedman, PhD2
1 Division of Pediatric Hematology/Oncology/Stem Cell Transplant, Columbia University Medical Center, NY; 2 Department of Biomedical Informatics, Columbia University, NY.

Introduction In the past, the high cost, high labor, and slow turn around time had prevented genotyping to directly impact patients at the point of care. Although we are rapidly approaching the reality of a $1000 genome and seeing breakthroughs in technology that can return exome-sequencing results in a matter of days,1 there are still great challenges in incorporating this genomic information into daily practice. For instance, there are currently no standards for the reporting of genetic variants discovered from genomic sequencing in clinical documents returned to the provider. This results in an inconsistent reporting practice across institutions.2

Our research goal was to represent the elements that are routinely found in genomic reports at our institution.

Methods This is a retrospective study of 100 genomic sequencing reports in the electronic health record (EHR) for patients performed at NYP/Columbia University Medical Center, NY from 2013-2014. Examples of test types include “Whole Exome Sequencing (WES),” “Columbia Combined Genetic Panel (CCGP),” “Cancer Whole Exome Sequencing (CWES),” etc. The corpus content was manually analyzed and an information model was created to adequately cover concepts described in these reports. This study was approved by the Institutional Review Board at our institution.

Results Variant reporting is very complex and the summary of the finding often has a many components. We have created a model to represent how genetic variants are reported at our institution (figure 1). For example, the variant reporting contains the gene name in where the variant is found, a variant frequency (proportion of the abnormal copies of the variant among all the normal copies of the gene, reported as a percentage), and a modifier. The modifier generally indicates whether or not the variant has previously been reported in the literature.

Conclusions Our model represents the elements described when reporting genetic variants at our institution. These variants are crux of the reports; they are what most clinicians are looking to read. By creating this model, our goal was to attempt to standardize reporting within our institution and add to existing efforts to standardize genomic variant reporting in sequencing reports returned to clinicians. Our model includes some common elements3 as reported by other institutions however there still needs to be a universal consensus. Future work necessary includes collaboration with outside institutions to increase the generalizability of our model.

References
An Exploratory Analysis of Inpatient Satisfaction and Usage Pattern of Personalized Bedside Station

Borim Ryu, MS, Seok Kim, MS, Kee-Hyuck Lee, MD, MS, Hee Hwang, MD, Ph.D.
Soyoung Yoo, Ph.D

1Center for Medical Informatics, Seoul National University Bundang Hospital, Seongnam, South Korea

Abstract

The Smart Bedside Station (SBS) is a screen that is installed on the bedside for the personal use and provides a variety of convenient services for the patients. This study aims to identify information needs, satisfaction degree and current system use status of SBS at the tertiary general university hospital. We examined inpatient response by satisfaction survey and system log data of each patient. The experimental result indicates the system importance and inpatient satisfaction.

Introduction

With increasing technological development, mobile devices and related applications have been modernized, and high-performing bedside terminal devices with numerous features have been introduced for hospitalized patients. The design of such devices takes into account the device performance, quality of contents, and user-friendly structure and is based on the user experience. The main purpose of this research is to (a) explore the end-user satisfaction in hospital information systems and (b) derive patient information needs in order to improve SBS system. Preliminary results can be expected to support patient-centered healthcare service.

Methods

The SBS terminal OS was equipped with Android version 2.3 and integrated with Electronic Health Record (EHR) system. The SBS system provided patients with several services and we grouped these services into three representative menus; 1) My Menu that supports laboratory test, prescription explanation, operation, meal, admission/discharge and doctor round, 2) Information Support which consists of education contents, message checking, medical expenses and hospital guide, and 3) Convenience Service such as meal ordering, nutrition information, satisfaction survey and other convenient service ordering.

We conducted direct user interviews, patient online survey and log analysis of bedside terminal system for a year. A total of 1,860 patient survey and 11,240 user log data were explored in this study, from December 2013 to 2014. All experimental method has been done in hospital settings and with multiple users simultaneously.

Results

With the experimental data analyses, we could derive system usage patterns by month/year and functional needs from the patients in clinical setting. Survey respondent relationship with the patient, age, gender, education status and degree of familiar with the smart device were examined. Especially, education status and degree of familiar with the smart device were statistically significant with the overall satisfaction. Above many SBS services, Information Support was the most frequently used menu except entertainment (TV/Internet) that provide general health information, message notice, hospital bills inquiries and hospital guide information. The experimental results suggest that there are many opportunities to improve the efficiency of information delivery and task performance to reduce system complexity, improve service quality and improve user satisfaction.

Conclusion

Recently, bedside terminal systems have been increasingly adopted in hospitals due to the rapid growth of advanced technology in healthcare. Further study should be directed at developing sophisticated patient service with the potential aspects of information seeking and usage patterns.

References

1. Miettinen S, Koivisto M. Designing services with innovative methods: University of Art and Design; 2009.
Demographic Predictors for Completion of an Interactive Voice Response System Survey Coupled with a Real Time Transfer to a Pharmacist

Alejandra Salazar, PharmD1; Elissa V. Klinger, MS1; Jeffrey Medoff, BS1, Mary Amato, PharmD, MPH1,3; Patricia Dykes PhD1, Jennifer S. Haas, MD1; David W. Bates, MD, MSc1,2; Gordon Schiff, MD1,2

1 Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA; 2 Harvard Medical School, Boston MA; 3 MCPHS University, Boston, MA

Introduction
In a previous study we demonstrated that an automated call-based pharmacovigilance system that integrates patient data from Electronic Health Records (EHRs) with an Interactive Voice Response System (IVRS) permitted primary care providers to identify drug related symptoms experienced by their patients (1). The purpose of this analysis is to study the demographic characteristics that predict a successful real time transfer to a pharmacist using an active pharmacovigilance system that incorporates data from the EHR into an IVRS.

Methods
A system was designed to collect patient demographics and medication related information from the EHR and upload them into an IVRS. Patients were invited to participate in an automated survey designed to screen for ADRs. At the end of the IVRS survey patients were transferred in real time to the study’s Clinical Pharmacist. The main demographic characteristics of interest for this analysis included age (18-30 years of age; 31-49 years; 50-74 years; and 75 years or older); sex; primary language (English or Spanish); and race as coded in the EHR (white, black, Hispanic/Latino).

Results
From the 11,945 eligible patients that were identified over 15 months, 5574 patients were randomized into the intervention. From that group, 4,847 patients received 5 to 10 IVRS call attempts and 3,301 patients (27.6%) answered the phone call. From those who successfully initiated the automated phone survey (609), 70.6% (n=430) patients were live transferred to the study pharmacist representing 13.03% of those contacted by the system.

Of the 430 patients who were transferred to the pharmacist to discuss their symptoms, 284 (66.04%) were female and 146 (33.95%) male (compared to 62% of females in the overall study population); 383 patients (89.07%) were English speaking and 41 (9.53%) Spanish-speaking. The racial distribution corresponded to: 45.12% White (n=194), 11.16% Black (n=48), 10% Hispanic, 1.16% Hispanic (n=5). Whites were most likely to finish the survey and progress to the real time transfer to the pharmacist (p=0.0079) compared to the other racial groups.

The age allocation of this cohort tallied: 59.77% (n=257 patients) were 50 to 74 years old, 22.09% (n=95) were 31-49, 11.16% (n=48) were 75 and older, and 6.98% (n=30) were 18 to 30 years old. In this category, the group most likely to be transferred to the study Clinical Pharmacist was the one corresponding to patients 50 to 74 years old (p=0.0095).

Discussion
We found that white race and age 50 to 74 were significant predictors of IVRS completion and real time transfer to the pharmacist for discussion of possible medication related problems. The highest overall rates of survey completion coupled with the live transfer were found for the group of English-speaking females in the aforementioned race and age category.

These results could have implications for the design and usability considerations of IVRS outreach calls in similar urban primary care settings.

This study was funded by grant # 5U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ).

References
A Multi-Dimensional Consumer-Oriented Approach to Evaluate Patient Portals

Mandana Salimi, BS, Adriana Stanley, BS, Vickie Nguyen, MS, Sahiti Myneni, PhD, MSE
The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX, USA

Introduction
Patient portals are used to promote well-being and facilitate communications between patients and their providers. Although mobile technologies and health options are growing rapidly, patient portals are still not well used. We investigate how patient portal applications, both on mobile and web platforms, currently function from a behavior change technique, persuasive design principle, and usability perspective while taking into account various Gestalt principles, such as simplicity, that improves the quality of the interface.

Research Methods
Two patient portal applications were evaluated. The patient portal mobile app of a mid-sized medical facility was downloaded from the Google Play store. The web-based app was also accessed through a website log in page. The link to the apps and websites were provided by the staff of the medical facilities. Our evaluation of behavior change techniques was based on those developed by Abraham et al. To evaluate whether or not the patient portal apps had successfully adhered to persuasive design principles suggested by the persuasive systems design (PSD) model, we mapped the functions, features, and content of the patient portal applications to each persuasive design principle. Two health informatics graduate students separately evaluated these two patient portal applications according to the above principles. After completing the mapping of functions, features, and content of the applications to these factors, the evaluators assessed their reliability of agreement. The evaluators had a reliability of $k=0.65$, indicating fairly strong consistency and reliability in our evaluations. To evaluate these apps based on the persuasive design principles, we took each category and divided the total number of principles that each patient portal was positively marked for by the total number of principles suggested for that category. Similarly, we evaluated the patient portal applications based on their adherence to the fourteen heuristics of usability guidelines and on Gestalt principles – principles that describe different ways in which we tend to visually perceive individual elements into groups.

Discussion and Conclusion
Patient portal applications can help patients better manage their health and maintain healthy behaviors. As shown in Figure 1, we found that none of the patient portal applications employed behavior change techniques and have an overall low adherence, 46%, to persuasive design elements. However, on the positive side, we found patient portals had an overall high percentage of conformity, 90%, to usability guidelines.

Figure 1. The patient portal adherence based on our three evaluation principles.

References
Changing Physician Changeover: How adopting a tool in the EMR impacts the perception of paper handoff tools

Jonathan D. Salisbury MD, Indiana University School of Medicine and Emily C. Webber MD, Indiana University School of Medicine and Riley Hospital for Children at IU Health

Background
Resident physicians change over care of their patients frequently. Poor communication during these handoffs is the most common cause of serious safety events in the hospital setting. Standardized handoff tools have been shown to improve resident handoffs and improve patient safety. Electronic medical records (EMR) have been widely adopted and have been identified as one way to optimize physician changeover. Paper handoff tools and checklists generated from EMRs are still frequently utilized by physicians, but have not been widely studied. The objective of this study is to identify the impact of the EMR implementation of a consistent, standardized handoff tool on resident changeover.

Methods
The study period began in October 2014. Investigators identified champions for the handoff project and presented the project at multiple faculty forums. A previously validated handoff methodology was adopted which focused on standard use of fewer fields, clear expectations, and increased situational awareness (a ‘watcher’ status). Resident physician education regarding the change consisted of a resident conference along with electronic versions of the educational materials. The change in the structure of the EMR handoff tool (Fig 1, Fig 2) was built, validated and implemented in December 2014. Impact of the implementation was assessed in two methods during this study. First, in the month preceding and immediately after the intervention, EMR generated resident changeover sheets were collected and assessed for added information and deleted information (strikethroughs), in addition to other study endpoints. Second, residents were surveyed electronically to gauge their perceptions of physician handoffs prior to implementation of standardized changeover.

Results
The Wilcoxon rank-sum test was used to compare the distributions of the additions per patient and strikethroughs per patient in the pre- and post-implementation conditions. The count of additions per patient is significantly lower after implementation than before implementation (p = 0.035). The count of strikethroughs per patient is significantly lower after implementation than before implementation (p = 0.005). For our second method of analysis, nonparametric Wilcoxon rank-sum tests were again used to compare resident perception of changeover. Residents were found to be significantly more likely to believe that they have adequate information (p=0.018) and that they were prepared (p=0.019) in the post-implementation period compared to the pre-implementation period. The overall length of the paper changeover tool per patient did not change significantly nor did the amount of time residents changed over face to face.

Discussion
Analysis of resident use of EMR generated changeover tools after implementing a standardized method for handoff was found to significantly decrease the amount of additional information residents add or delete. We believe that the consistency of a standard handoff tool, as well as more limited, well delineated fields in the guide, helped improve the accuracy and completeness of both the verbal and paper changeover tools. At the conclusion of the study period, we identified ways to optimize resident physician changeover and improve patient safety in the hospital setting.

Figure 1. Pre-intervention sign off guide generated by EMR (5 fields for physicians to edit).

Figure 2. Post-intervention sign off guide generated by EMR (3 fields for physicians to edit, plus status and synthesis question).
MAC Annotator: An interactive tool for translating medication appropriateness criteria into structured form

Hojjat Salmasian, Carol Friedman, PhD

Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

Inappropriate use of medications (IUM) is a global problem leading not only to excessive cost of care, but also putting patients at risk of avoidable harm. Traditionally, to identify IUM, clinicians use a set of guidelines, commonly referred to as Medication Appropriateness Criteria (MAC). Previously, we developed a framework for structured representation of MAC.1 To facilitate translation of MAC into structured form, we developed and validated a tool we call MAC Annotator, an interactive tool for annotation of MAC.

Methods

MAC Annotator is a client-side web-based program using the jQuery library. It allows the user to highlight a part of the narrative string, and to associate it with a concept in an existing biomedical terminology. It interacts with the NCBO Annotator application programming interface (API). It first uses the API to “pre-annotate” the MAC by identifying mentions of medication names and disease names using three terminologies (MeSH, RxNorm and SNOMED CT). It then allows users to add, modify or delete annotations, and facilitates this by providing a straightforward search mechanism for all terminologies available. We measured how many of the concepts could be automatically identified using the pre-annotation, how many could be found only by the user, and how many had to be defined manually (i.e. could not be found in any existing terminology). We also measured the time spent for each of these activities. To measure the agreement between annotators, we divided twice the number of concepts that were identically annotated by both raters to the total number of concepts identified by both annotators. This proportion can range between 0 to 100%, where higher numbers indicate more agreement between annotations.

Results

Six experts each annotated six MAC using MAC Annotator. The average time spent to complete the task was 121 seconds, of which 0.50 was spent for automated pre-annotation. Pre-annotation resulted in identification of 75 concepts, from which experts modified seven (9%) and accepted the remainder as correct. In the end, 180 concepts were annotated in the MACs, of which 107 referenced a concept in a terminology while the remainder were manually defined by the experts. The agreement between annotations was 77.3%. Agreement between annotators was 100% for medication names, and 100% for medication class names. With disease entities, agreement was only 50%. The main reason for disagreement was that annotators used different approaches when the disease was mentioned along with a modifier. For instance, one annotator broke down New York Heart Association Class III or IV heart failure to several concepts but the other used a manual definition for the entire term (in both cases, no terminology contained an equivalent term hence definitions were manual). Also, one reviewer provided a manual definition for the term uncomplicated pulmonary embolus while the other mapped pulmonary embolus to an existing terminology and provided an explanation for the term uncomplicated. The definition of certain terms was context-dependent. Examples include uncomplicated (as a modifier before a disease name) or monotherapy (which can be translated as “excluding other therapies for the same disease”), where the other therapies change for each disease. While some annotators provided context-specific definitions, others provided abstract definitions (e.g. defining uncomplicated as “with no complication”) that would be context-agnostic but not objective or computable.

Discussion

MAC Annotator facilitates the annotation of MAC through the pre-annotation step as well as through providing an easy to use, interactive interface for searching existing terminologies. It also helped identify content that is not well-defined in the original MAC, which can lead to different interpretations. Traditionally, annotation of biomedical text is done by a small number of expert annotators over extended periods of time.2 Previous work on creating structured versions of decision support rules also relies on manual annotation.3 MAC Annotator may help this process by speeding up the translation of these narratives into structured form. Pérez-Pérez et al have recently conducted a thorough review of public and well-known annotation tools. Compared to existing tools, MAC Annotator has the advantage of pre-annotating the text in a time-efficient manner with high accuracy, allowing annotators to define terms that are not explicit, and platform-agnostic design which doesn’t require any software installation by the user.

References

Integrative Informatics and Predictive Modeling Support for Population Health

Mary M. Saltz, MD, Joel H. Saltz, MD, PhD, Janos Hajagos, PhD, Andrew White, PhD, Charles Boicey, MS, RN-BC, Jim Murry BS, Ivan Crnosija, BA, Tahsin Kurc, PhD, Erich Bremer, MS, Jonas Almeida, PhD
Department of Biomedical Informatics Stony Brook School of Medicine, Stony Brook, NY

Introduction
This project targets development of proactive population health informatics methods in a geographically distributed medically and culturally heterogeneous population. The informatics infrastructure will be generally applicable to population health clinical informatics efforts, and is being developed in the context of the Center for Medicare and Medicaid (CMS) funded, 450,000 life (Medicaid and Uninsured), 3000+ organization New York Suffolk County DSRIP Performing Provider System (PPS). Future payments to the PPS will be directly tied to the overarching metrics of 25% reduction in potentially preventable hospitalizations (PPR) and emergency department visits(PPV) [3M Health Information Systems] and to improvements in patient quality indicators such as Agency for Healthcare Research and Quality (AHRQ) Prevention Quality Indicators (PQI) and Pediatric Quality Indicators (PDI) metrics. The complex interactions of each stand-alone project and the overall DSRIP metrics is such that changes to any part of this ecosystem will impact others. To this end, it is not enough to merely identify outcomes and trends, but also to design an in silico test bed to evaluate and thus steer proposed interventions or changes to the system.

Methods and Objectives
In order to plan, coordinate and steer population health activities, we need to assemble a 360 degree view that: 1) encompasses patient health, co-morbidities, geographic location; 2) takes into account each provider’s patient population, 3) is able to attribute PPRs, PPVs, PQIs and PDIs to health care systems and to providers. Creation of 360 degree health care views requires us to integrate multiple site electronic health records, New York State data from Statewide Planning and Research Cooperative System(SPARCS) and public health data including US Census Bureau, Behavioral Risk Factor Surveillance System(BRFSS) and American Community Survey data, along with data from deployed smart phones, mobile sensors such as Fitbits, Apple Watch and instrumented home medical devices.

Data integration pipelines, data quality control pipelines, data dictionary and data products are being created using a variety of tools, including a novel open source, streaming JavaScript based OpenHealth platform (https://github.com/mathbiol/openHealth). The tools are designed and configured to take secure advantage of a combination of open and protected data sources. Data products encompass clinical/computational phenotype generation - e.g. validated co-morbidities, treatment history and demographics. In the near future, data will also characterize patients by provider with respect to quality indicators such as PPRs and PPVs. Data will also include computed quality indicators, detailed insight into demographics using geo-located American Community Survey/Census data along with project specific data marts.

Success in the core DSRIP objectives of 25% reduction in PPRs and PPVs require systematic targeting of interventions and hence require coordinated population and risk stratification, as well as iterative adaptive analyses to allow on the fly interventions. Effective efforts to improve care transitions require projects to identify patients at risk and to characterize patients specific risks. This project will leverage a variety of machine learning methods including the techniques reported in [1] as data is collected as the DSRIP program is implemented.

Conclusion
While this project is still in its early stages, our tools and methods have already played a pivotal role in generating the detailed population health characterization that led to our near perfect Community Needs Assessment NY State DSRIP proposal score. The tools and methods developed have laid a foundation for predictive population health analytics planned over the 5 years of the project.

References
Predicting Acute Kidney Injury in Critically Ill Children Using Electronic Health Record Data – A Comparison of Four Statistical Learning Models

L. Nelson Sanchez-Pinto, MD1; Robinder G. Khemani, MD MsCI1. 1Pediatric Critical Care Medicine, Children’s Hospital Los Angeles, Los Angeles, CA

Background
Development or progression of acute kidney injury (AKI) is associated with increased mortality in the pediatric intensive care unit (ICU). Serum creatinine, the main AKI biomarker, is a late marker of the disease. Clinical prediction models can potentially identify high-risk patients earlier in their disease course. Our goal was to develop and compare four clinical prediction models of AKI in the pediatric ICU using electronic health record (EHR) data.

Design/Methods
We analyzed all patients 1 month to 21 years of age with no chronic kidney disease admitted to a single pediatric ICU between 3/2003 and 5/2012, who were in the ICU at least 24h. AKI was determined using the Kidney Disease Improving Global Outcome (KDIGO) criteria. Potential risk factors for AKI were analyzed using EHR data from the first 12 hours of ICU care and were compared between those with new or worsening AKI in the first 72 hours of ICU admission versus those with no AKI or no AKI progression. Risk factors were based on the pathophysiology of AKI with a specific emphasis on variables that are objectively recorded in the EHR and generalizable across pediatric ICUs. The dataset was randomly divided into training and validation sets (60/40 split of patients) and four statistical learning models were applied to the training data: logistic regression, classification tree, random forest, and support vector machine. Models were compared based on the area under the curve (AUC) of the ROC plot, sensitivity, specificity, positive and negative predictive value and likelihood ratios on the validation set.

Results
7029 patients met inclusion criteria, 6.3% (440 patients) developed new or worsening AKI within 72 hours of admission and these had higher mortality than those without AKI (28.9% vs. 3.5%, p<0.001). The original dataset contained 90 variables but was reduced to 8 variables for the final based on multivariate association with the outcome and the objective/generalizable criteria. The variables used were: age, lowest pH, lowest platelet count, highest blood urea nitrogen level, first serum creatinine level, highest bilirubin level, and whether the patient had a cardiac arrest or an operative procedure prior to the ICU admission. The performance of the four statistical learning models using these 8 variables is shown in Table 1. All models had the best performance with the same set of clinical variables. All models had a strong NPV and negative LR. Random Forest had the highest positive LR.

Table 1. Performance of the four statistical learning models in the validation set (n=2829, AKI+=179).

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>Classification Tree</th>
<th>Random Forest</th>
<th>Support Vector Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test+ (%)</td>
<td>703 (25)</td>
<td>845 (30)</td>
<td>617 (22)</td>
<td>684 (24)</td>
</tr>
<tr>
<td>Sensitivity % (95%CI)</td>
<td>80 (74-86)</td>
<td>79 (72-84)</td>
<td>80 (74-86)</td>
<td>78 (72-84)</td>
</tr>
<tr>
<td>Specificity % (95%CI)</td>
<td>79 (79-79)</td>
<td>73 (73-74)</td>
<td>82 (82-83)</td>
<td>80 (79-80)</td>
</tr>
<tr>
<td>PPV % (95%CI)</td>
<td>21 (19-22)</td>
<td>17 (15-18)</td>
<td>23 (22-25)</td>
<td>21 (19-22)</td>
</tr>
<tr>
<td>NPV % (95%CI)</td>
<td>98 (98-99)</td>
<td>98 (98-99)</td>
<td>98 (98-99)</td>
<td>98 (98-99)</td>
</tr>
<tr>
<td>LR+ (95%CI)</td>
<td>3.8 (3.4-4.1)</td>
<td>3.2 (2.7-3.2)</td>
<td>4.5 (4.4-4.9)</td>
<td>3.8 (3.4-4.2)</td>
</tr>
<tr>
<td>LR- (95%CI)</td>
<td>0.3 (0.2-0.3)</td>
<td>0.3 (0.2-0.4)</td>
<td>0.2 (0.8-0.3)</td>
<td>0.3 (0.2-0.36)</td>
</tr>
<tr>
<td>AUC (95%CI)</td>
<td>0.80 (0.77-0.83)</td>
<td>0.76 (0.73-0.79)</td>
<td>0.81 (0.78-0.84)</td>
<td>0.79 (0.76-0.82)</td>
</tr>
</tbody>
</table>

PPV, positive predictive value; NPV, negative predictive value; LR, likelihood ratio.

Conclusions
We developed, validated and compared four objective clinical prediction models of AKI development and progression in the first 72 hours of ICU admission using variables that are easily extractable from EHRs and generalizable across pediatric ICUs. There were slight differences in performance between models, but they all had strong NPV and negative LR. Random forest had the best positive LR amongst the models. Clinical prediction models such as these can be used to develop clinical decision support tools to risk stratify patients, flag patients that could benefit from a urine AKI biomarker test, implement preventive bundles, or design clinical research studies.
Canary – a Graphic User Interface to a Heuristic NLP Engine

Nicolae L. Sandor, DM1, Stephen Skentzos, MA2, Alexander Turchin, MD, MS3,4
1Boston University, Boston, MA; 2Kanjoya, Inc, San Francisco, CA; 3Brigham and Women's Hospital, Boston, MA; 4Harvard Medical School, Boston, MA

Introduction
A large fraction of the Big Data in the EMR is difficult to analyze because of its narrative format. Narrative data can be studied using NLP techniques. However, many available NLP packages are either very costly or require that their users have substantial programming skills, limiting their adoption by mainstream clinical researchers.

We previously developed a versatile NLP engine that could implement both context-free and context-dependent grammar1 that was successfully used in large-scale analyses2. We have now developed Canary - a graphic user interface (GUI) to this engine that allows its use by individuals without access to software development resources.

Materials and Methods
The NLP engine that Canary uses allows the user to define a language model using the following components: a) word classes (ontology); b) phrase structures (grammar rules); c) order in which phrase structures are used; d) goal state(s); and e) output(s). It takes as input one or multiple text files. The engine is implemented in Perl. It supports multi-threading and hierarchical processing, attaining speeds of 40 documents/second.

Results
Canary implements a GUI for the NLP engine as a .NET based standalone Windows executable. Canary allows the user to enter / edit all of the language model components and apply the language model to a selected dataset to generate output over 11 dialog screens (See Figure 1 for an example). Canary interface assumes no knowledge of software development or NLP techniques. Canary executable and code are available upon request free of charge.

![Canary Interface for Entry of Word Classes](image)

Figure 1. Canary interface for entry of word classes (custom ontology).

References
Meta-Analysis: Impact of Health Information Technology on Patient Engagement and Health Behavior Change

Suhila Sawesi, MS, PhD; Kanitha Phalakornkule, MS, PhD; Josette Jones, RN, PhD

\(^1\)IUPUI, Indianapolis, IN, USA

Abstract
This poster presents a meta-analysis of the impact of information technologies (IT) on patient engagement and behavior change. The primary goals of this review were to systematically determine, (1) impact of information technology (IT) platforms used to promote patients’ engagement and to effect in change health, (2) behavioral theories applied as bases for developing these interventions and their impact on health outcomes, (3) different ways of measuring health outcomes, (4) usability and acceptability of these technologies, (5) challenges and research directions for implementing IT platforms to meaningfully impact patient engagement and health outcomes.

Background
To reach optimal benefit from treatment, patients must engage in managing their health, not just simply following a physician’s recommendation.[1] Frequent, real-time communication and feedback are essential in supporting health behavior change and empowering patients’ engagement in the health care process. [2] However, the traditional model of care delivery, a face-to-face interaction with an expert and/or trusted health care provider, can be implemented only with a small number of patients and thus has limited impact. [2] In an effort to reach larger number of patients, researchers and clinicians have begun exploring the role of information technology (IT) platforms in patient engagement and health behavior change interventions.[2-3] IT interventions generally have potential benefits and some proven effects; however, specific components of the IT interventions associated with their success remain unclear. To better understand how to build successful IT interventions that can change patient behavior in a meaningful way, a meta-analysis was performed.

Method and result
PubMed, Web of Science, PsycINFO, and Google Scholar were searched for studies published between 2000-December 2014. Two reviewers assessed the quality of included papers and potentially relevant studies were retrieved and assessed for eligibility based on predetermined inclusion criteria. A total of 170 articles met the inclusion criteria and were reviewed in detail. Overall, 89% (151/170) of studies showed positive impact on patient behavior and 83% (141/170) reported high levels of improvement in patient engagement. Only 47% (80/170) referenced specific behavior theories and only 34% (57/170) assessed the usability of IT platforms. The majority of studies used indirect ways to measure health outcomes (66%, 112/170). Moreover, the follow-up assessments ranged from extremely short durations (1 week) to long-term ones (36 months), making it unclear as to how long the effects of these technologies would last. This review found no significant relationship between the studies that implemented the theory of behavior change and the impact of technology (P=0.97). Also, this review found no significant association between ways to measure health outcomes and the impact of technology (P=0.446). Lastly, the result showed significant association between usability and patients’ engagement in health care (P=.0216).

Conclusion
In general, the review has shown that IT platforms can enhance patients’ engagement and improve health outcomes. Few studies addressed usability of these interventions. Also, the reason of not using specific behavior theories remains unclear, and further research is needed to clarify these important questions. In addition, an assessment these types of interventions should be conducted based on a common framework using a large variety of measurements; these measurements should include those relating to motivation for health behavior change, longstanding adherence, expenditure, satisfaction, and health outcomes.

References
Automating risk score calculations and care recommendations by an EMR agnostic solution and potential time saving for providers

Marianne Scheitel; Hongfang Liu PhD; Jane L. Shellum; Ronald Hankey M.S. MBA; Steve Peters MD; Rajeev Chaudhry MBBS MPH
Mayo Clinic, Rochester, MN

Abstract

MayoExpertAdvisor (MEA) provides patient-specific risk score calculations and care recommendations using data from electronic medical records (EMRs). We report a timing study that evaluates the baseline cognitive effort and accuracy for the calculation of risk stratifications and recommendations for chronic heart failure (CHF), atrial fibrillation, and atherosclerotic cardiovascular disease (ASCVD). The study demonstrates that MEA can save time for providers in delivering automated risk score calculations and care recommendations with a great potential to improve compliance with recommended treatments.

Introduction

We developed MayoExpertAdvisor (MEA) to deliver care recommendations based on care process models (CPMs) established at Mayo Clinic based on national guidelines and expert consensus. MEA can calculate risk scores using data from electronic medical records (EMRs) with the goal of saving time and promoting consistency of care. The purpose of this study is to establish baseline data for efficiency and accuracy and to assess the potential for improvement.

Methods and Results

A fixed patient scenario was given to 22 primary care physicians, 4 residents, and 4 nurse practitioners. Each provider was asked to recommend treatment for CHF, atrial fibrillation, and hyperlipidemia and to calculate the associated risk scores (Seattle Heart Failure Model (SHFM)¹, CHA²DS²-VASc², and American College of Cardiology ASCVD risk respectively³). Each provider was recorded and analyzed with Morae ® (version 3.3.3) software⁴. The efficiency metrics collected includes the time and the number of clicks, keystrokes, and web page changes for each recommendation. The accuracy was determined by the number of providers that gave the correct risk scores and recommended care based on the corresponding CPMs.

The efficiency metrics are shown in Table 1. The CHA²DS²-VASc and ASCVD risk score were calculated with 90% accuracy. The capability of MEA calculating risk scores and recommending treatments automatically has the great potential to save significant time for providers and to improve accuracy of treatment.

Table 1: Mean Efficiency Values (Standard Deviation)

<table>
<thead>
<tr>
<th>Care Recommendation Tasks</th>
<th>Elapsed Time (MM:SS)</th>
<th># Clicks</th>
<th># Keystrokes</th>
<th># Page Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHFM + CHF CPM</td>
<td>5:47 (1:29)</td>
<td>159 (60)</td>
<td>27 (11)</td>
<td>8 (3)</td>
</tr>
<tr>
<td>CHA²DS²-VASc +Atrial Fibrillation CPM</td>
<td>1:25 (0:17)</td>
<td>55 (24)</td>
<td>13 (7)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>ASCVD Risk + Lipid CPM</td>
<td>1:50 (0:31)</td>
<td>70 (25)</td>
<td>33 (18)</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Total</td>
<td>9:02 (1:34)</td>
<td>284 (84)</td>
<td>73 (28)</td>
<td>15 (4)</td>
</tr>
</tbody>
</table>

Conclusion

The capability of MEA calculating risk scores and recommending treatments automatically has the great potential to save significant time for providers and to improve accuracy of treatment.

References

Statistically Bolstered Opportunities Assessment in Measure Analytics

Elisabeth Lee Scheufele, MD, MS1, 2, Julianna Kohler, MHS1, Gerardo Soto-Campos, PhD1
1ConvergeHealth by Deloitte, Newton, MA; 2Harvard Medical School, Boston, MA

Abstract

We describe the use of the binomial test of proportions as a means to calculate the impact of patient level attributes on a quality or performance metric's population with a measure of statistical significance. Hospital administrators can use the output of this methodology as a statistically bolstered opportunities assessment and gain insight into potential detrimental or protective attributes of a metric's populations for investigation.

Introduction

Health care systems are actively working on multiple simultaneous performance and quality improvement initiatives. Typically, analysts monitor these activities via measures and metrics that calculate rates and averages and then focus operational efforts on those metrics that are performing outside of expectations. Investigating the metric's population for opportunities for improvement or for subgroups that demonstrate beneficial tendencies is often a particularly manual process. We introduce a methodology that allows statistically bolstered insight into patient attributes that may impact the metrics of interest to the health care system.

Methodology

We applied the binomial tests of proportions 1 to identify patients' descriptive attributes that may provide insight into possible beneficial or detrimental impact on measure scores. We apply this methodology to metrics numerically defined as a ratio with a denominator definition and a numerator that is defined as a subset of the denominator population. With this metric structure, the denominators of the two proportions are defined as the group of patients who make up the numerator (“meet the numerator definition”) and the group of patients who make up the remainder of the denominator (“not meet the numerator definition”). Then within these two proportion denominators, the proportion of patients with certain patient attributes of interest (e.g., “Caucasian”, “diabetes”, “18-44years - Age bucket”, “PCP: Dr. Smith”, etc.) are calculated. The binomial test of proportions is applied to the two proportions, producing a z score with a p-value. In Table 1, we demonstrate, with simulated values, the potential outcomes and interpretation of applying this methodology to a metric. In this example, the metric is one where maximizing the rate is optimal (e.g., “Rate of patients who have had influenza vaccinations”). It is necessary to make this distinction because it affects the applicability of the potential interpretation. In this case, the higher proportion in those who meet the measure could indicate a protective or beneficial attribute, while a higher proportion in those who do not meet the measure would indicate a possible detrimental or high risk attribute.

<table>
<thead>
<tr>
<th>Attribute Type</th>
<th># with Attribute (n = 1000)</th>
<th>% w/ Attribute</th>
<th>Meet Num Definition</th>
<th>Not Meet Num Definition</th>
<th>p-value</th>
<th>Meet/Not/ NS</th>
<th>Potential Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race #1</td>
<td>460</td>
<td>46%</td>
<td>46% 465 750</td>
<td>46% 116 250</td>
<td>0.0232</td>
<td>Not</td>
<td>NS</td>
</tr>
<tr>
<td>Race #2</td>
<td>155</td>
<td>16%</td>
<td>14% 105 750</td>
<td>20% 50 250</td>
<td>0.4226</td>
<td>NS</td>
<td>Not Detrimental</td>
</tr>
<tr>
<td>Race #3</td>
<td>353</td>
<td>35%</td>
<td>36% 270 750</td>
<td>33% 83 250</td>
<td>0.0319</td>
<td>Meet</td>
<td>Protective</td>
</tr>
<tr>
<td>Practitioner #1</td>
<td>113</td>
<td>11%</td>
<td>10% 75 750</td>
<td>15% 38 250</td>
<td>0.0245</td>
<td>Not</td>
<td>Detrimental</td>
</tr>
<tr>
<td>Chronic condition #1</td>
<td>123</td>
<td>12%</td>
<td>12% 90 750</td>
<td>13% 33 250</td>
<td>0.6171</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Chronic condition #2</td>
<td>333</td>
<td>33%</td>
<td>31% 233 750</td>
<td>40% 100 250</td>
<td>0.0054</td>
<td>Not</td>
<td>Detrimental</td>
</tr>
<tr>
<td>Chronic condition #3</td>
<td>333</td>
<td>33%</td>
<td>31% 233 750</td>
<td>40% 100 250</td>
<td>0.0054</td>
<td>Not</td>
<td>Detrimental</td>
</tr>
<tr>
<td>Age Bucket #1</td>
<td>333</td>
<td>32%</td>
<td>32% 240 750</td>
<td>33% 83 250</td>
<td>0.7256</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

Table 1. Simulation of applying the binomial test of proportions on a maximizing metric

Conclusion

We have applied a statistical methodology to potentially streamline the work of identifying patient attributes that could be considered detrimental or protective to a measure of interest for a health care system. Primary limitation is that the binomial test of proportions is based on a z-score argument1, which assumes normality. Its applicability weakens as N becomes smaller. For the case of low N, we are considering bootstrap approaches2 to generate significance criteria.

Reference

Clinical Relevance of the Doctor’s Dilemma Question Set

Daniel R. Schlegel, Sashank Kaushik, Peter L. Elkin
Department of Biomedical Informatics, University at Buffalo, Buffalo, NY

Introduction

Every year at the annual ACP meeting, residents compete in the Doctor’s Dilemma (DD) competition [1] — essentially medical Jeopardy. The asked questions cover topics in all of medicine and range from simple trivia to complex diagnosis and treatment decision questions. Questions are split into 26 categories including topic areas such as History and Poisoning; and entire medical subfields such as Oncology and Cardiology.

Many computational systems attempt to answer clinical questions. For example, there are a great number of decision support tools which make use of patient data and answer a specific set of questions. There are fewer systems which attempt to answer all types of clinical questions, our eventual goal. In developing such a system, it is important to have a set of questions to train on. Such questions must be: topically diverse, covering many different topics and scenarios in medicine; and clinically relevant — questions should be those which could occur to a doctor during clinical practice. Questions may be those a doctor should know the answer to, or they may be ones normally looked up. Even a cursory glance shows that the DD questions cover many topics and scenarios, but it is unclear how many are clinically relevant — the subject of this study. Only the IBM Watson team has previously made use of these questions in developing a question answering system. They only used diagnosis questions, which are clearly clinically relevant [2].

Methodology

The DD question set was obtained through the ACP website and communication with the ACP. As of February 2015 it consists of 1110 questions. Of those, 171 refer to images and were excluded since we are only interested in textual questions. The remaining 939 questions were divided randomly into two datasets — with one to be used in later projects and not examined by our group during this study. From the remaining 465 questions in our dataset, we attempted to extract at least 10 questions from each of the 26 categories for manual review. This was not always possible since at least 10 questions were not available for each category. The result was a sample of 229 questions. Two clinicians, SK and PLE, examined and annotated each question as either clinically relevant, or not. The two annotators were in near–perfect agreement (κ = .87). The disagreements were mediated and a consensus was formed.

Results

Very little of the DD question set was found to be not clinically relevant — only 8 of the 229 questions (3.5%) in our sample. The primary category where irrelevant questions occur is History. Out of the 10 questions from that category, the annotators found that 5 were not relevant clinically. Other irrelevant questions came from the Ethics category (1 of 6 found irrelevant), Biostatistics (1 of 6), and Epidemiology (1 of 10).

Discussion

Once the results were tabulated, we analyzed the question categories which contained questions which were not clinically relevant in more detail. We found that the History section tends to contain a large amount of trivia. Questions such as “Nobel awardee for discovery of insulin” were labeled irrelevant since they aren’t helpful in clinical practice, while others such as “Previous name for reactive arthritis” may be relevant to a clinician, even though it is historic data. From examining the Ethics and Biostatistics categories, it was easy to see that they may contain non-clinical questions such as those about IRB approval and certain equation usage. It was surprising to see a question from Epidemiology in the irrelevant category, but the question: “Greek letter that represents the probability of making a type I error”, would probably have been better categorized as Biostatistics. Since the DD questions are mostly clinically relevant, we believe the set is a good tool for the development of clinical question answering systems.

References


1We would like to thank the ACP for allowing us to use the Doctor’s Dilemma questions in this research.
Use of mHealth Technology for Supporting Symptom Management in Underserved Persons Living with HIV (PLWH)

Rebecca Schnall, PhD, RN¹, Haomiao Jia, PhD¹,², Susan Olender, MD⁴, Suzanne Bakken, PhD, RN¹,³
¹ School of Nursing ² Department of Biostatistics ³Department of Biomedical Informatics
⁴College of Physicians and Surgeons, Columbia University, New York, NY, USA

Abstract

The ability to self-manage adverse symptoms of HIV has been shown to improve patient-centered outcome in persons living with the disease. Mobile Health (mHealth) offers an ideal platform for the implementation and dissemination of evidence-based strategies for HIV symptom management. In our proposed work, we will develop a mHealth tool to provide real-time access to important symptom management strategies to improve patient-centered outcomes.

Introduction

HIV has evolved from an acute illness to a chronic illness¹ and patients, especially those with co-morbid conditions, have multiple adverse symptoms that require ongoing management. An individual's ability to self-manage the adverse symptoms of his/her HIV illness has been shown to improve patient-centered outcomes². In response to this need, a team at UCSF developed and demonstrated the efficacy of a paper-based symptom management manual with self-management strategies for common HIV/AIDS adverse symptoms³. However, subsequent use of these strategies has been very limited. To facilitate uptake, our study team developed and pilot tested a web-based system that delivers patient-centered tailored symptom management strategies for persons living with HIV (PLWH) to better manage adverse symptoms and improve overall quality of life⁴.

Approach

In our upcoming work, we will develop and evaluate the efficacy of incorporating symptom management strategies into a mobile health (mHealth) tool to improve symptom self-management for PLWH. Specific aims are to: 1) Apply participatory design methods to incorporate HIV symptom self-management strategies into a mHealth tool for use in patient self-management and 2) Using a randomized design, examine the effect of the mHealth tool as compared to a control group on primary outcomes of symptom frequency and intensity.

In this poster presentation, we will present findings from our participatory design sessions. Participants in our design sessions will include HIV patients and clinicians who will design the mHealth tool through user-centered design processes including use case. The resulting use cases will describe the behaviors of the user of the system and of the system itself and guide refinement of mHealth tool prior to usability testing.

Discussion

The ubiquitous nature of mobile technologies in daily life has created opportunities for health-related applications that were not previously possible. As such mHealth technology has the potential to be a highly valuable tool in the management of chronic illnesses such as HIV. Despite the growing use of mHealth technologies, many applications have not been rigorously evaluated and there remains limited evidence of their acceptability to intended users and impact on health care outcomes. Findings from this work may have broader implications for understanding if mHealth can improve efficiency and effectiveness of care for persons living with HIV.

References

Computerized Provider Order Entry Rates and Length of Stay Are Inversely Correlated

Richard Schreiber, MD, FACP1, Steven H. Shaha, PhD, DBA2
1Chief Medical Informatics Officer, Holy Spirit Hospital, Camp Hill, Pennsylvania—A Geisinger Affiliate; 2Professor, Center for Policy & Public Administration, Principal Outcomes Consultant, Allscripts

Intervention

This was a retrospective, correlational, interrupted time series analysis of CPOE rates and LOS measured over 30 consecutive quarters from July, 2007 through December, 2014. All patients and all orders were eligible except for medication orders in the ED which were not available electronically during the first part of the study from July, 2007 through December, 2012. During this interval there were 76,972 discharges and 6,135,994 eligible orders. All patients and all orders now including ED medications were reassessed similarly for the 8 quarters from January, 2013 through December, 2014, involving 25,795 discharges and 2,550,289 eligible orders. Case mix index (CMI) was tracked across all quarters. Statistical analyses included regression analysis, computation of Pearson’s correlation coefficients with arcsine, probit, and Fourier transformations, and calculation of R² coefficients and p values. Inflection point analysis was performed⁴.

Results

CPOE increases from 2008-2012 were significantly correlated with LOS reductions for 13 disciplines and organization-wide. Monthly CPOE-LOS data showed a linear, negatively-sloped relationship between CPOE and LOS (r = -0.798, p<0.001). The relationship was strongest as a sigmoidal curve (R² = 0.887, p<0.001). The inflection point⁴ where CPOE gained lasting favorable effects on LOS was at 60% adoption. Overall there was a 20.2% reduction in LOS attributable to CPOE. During 2013-2014 CPOE and LOS in seven disciplines that showed variation in CPOE rates exhibited converse, curvilinear trends precisely complementary to each other with R² of 0.76 and 0.73 respectively (p<0.001). Case mix index remained steady or rose from 2008 to 2012, and showed no change during 2013-2014 (t=0.036, p=0.81).

Conclusions

This study uses new methodology: it measured the impact of CPOE on LOS on a per-patient, per-visit, per-discipline basis thereby controlling for inherent differences in the rate of CPOE adoption by various providers. The authors believe this is a more accurate representation than overall summary statistics. This technique also helps to control for other efforts simultaneously undertaken to reduce LOS. This supports the hypothesis that CPOE contributes significantly to LOS, independently of the other numerous on-going efforts to reduce LOS.

The return on investment may not be apparent until CPOE rates exceed a tipping point. Attaining meaningful use stage 2 is necessary but not sufficient to realizing a return on investment of CPOE as measured by LOS.

References


There was no funding for this research. Dr. Shaha is a consultant to Allscripts, the EMR at this hospital, which was not involved in the concept, design, analysis, interpretation, manuscript writing, or decision to submit for publication. We thank Kevin Peters for his contributions.
Drug Allergy Interaction Alert Overrides in the Inpatient Setting

Diane L Seger RPh1,2, Sarah P. Slight MPharm PhD PGDip1,3, Patrick E Beeler MD1,9, Olivia Dalleur MPharm PhD1,4,5, Mary G Amato PharmD MPH1,6, Tewodros Eguale MD PhD1,6,7, Karen C Nanji MD MPH8, Patricia C Dykes PhD RN1,9, Michael Swerdloff BS1,

Julie M Fiskio BS2, David W. Bates MD MSc1,9

1Brigham and Women's Hospital, Boston, MA, USA; 2Partners Healthcare, Wellesley, Boston, MA, USA; 3School of Medicines, Pharmacy and Health, Durham University, Durham, UK; 4Louvain Drug Research Institute, Universite catholique de Louvain (UCL), Belgium; 5Cliniques universitaires Saint-Luc (UCL), Brussels, Belgium; 6MCPHS University, Boston, MA; 7McGill University, Montreal, Quebec, Canada; 8Massachusetts General Hospital, Boston, MA, USA; 9Harvard Medical School, Boston, MA, USA.

Abstract: Drug allergy interaction (DAI) alerts are often overridden in the outpatient setting. We evaluated the rate of DAI overrides in the inpatient setting and why providers chose to override these alerts. The override rate of DAI alerts in inpatients was found to be very high (83.9%). The total number of alerts fired for true allergic reactions (immune-mediated) was only 38,224 (29%) of which 31,191 (81.6%) were overridden.

Introduction: Clinical Decision Support (CDS) systems can provide clinicians with relevant, treatment-related information and improve patient safety at the point of prescribing. Drug allergy interaction alerts (DAIs) are generated when a drug is prescribed to a patient that has a known adverse reaction. However, high DAI override rates have been reported in the outpatient setting.(1) We evaluated the DAI override rates in the inpatient setting and the reasons why providers chose to override these alerts.

Methods: All inpatient DAI alerts from Jan 2009 to Dec 2011 were obtained from the Brigham and Women’s Hospital, Boston, MA. These alerts suggest an undesirable interaction likely to injure the patient, and give the provider the option of either ‘cancelling’ the order or ‘overriding’ the alert. A total of 2,783 prescribers received allergy alerts at this site. The downloaded file included the name of the drug and allergen that triggered the DAI alert, and the reasons that the providers gave at the time of overriding the alert. Duplicates were removed and replaced, and patient information anonymized prior to analysis.

Results: The rate of DAI overrides was 83.9% (110,414 /131,615). Opioids triggered more than half of the alerts (50.9%, n= 66,949), with only 7,358 (10.9%) of these accepted by providers. The total number of alerts fired for true allergic reactions (immune-mediated) was only 38,224 (29%) of which 31,191 (81.6%) were overridden. Surprisingly, providers overrode 75.1% (4,917/6,546) of even alerts that indicated that the patient was at risk of developing anaphylaxis or angioedema. The most common coded reasons for overriding these particular DAI alerts were ‘Patient has taken previously without allergic reaction/patient has tolerated previously’ (55.5%, 2,730/6,546), ‘Physician aware’ (13.8%, n= 678) and ‘Low risk cross sensitivity, will monitor’ (11.8%, n= 582). Drug classes which were more frequently associated with allergies triggered 31,843 DAI alerts (24.2%); leading classes were cephalosporins (13.7%, n= 17,978), aspirin and NSAIDs (6.6%, n=8,730), penicillins (1.8%, n= 2,379), sulfonamides (1.4%, n= 1,819), ACE inhibitors and angiotensin receptor blockers (0.55%, n=724), contrast media (0.12%, n=161), and monoclonal antibodies (0.04%, n=52).

Conclusion: The override rate of DAI alerts in inpatients was found to be very high. This study offers important insights into the specific DAI alerts that were commonly overridden and the reasons given by providers. Of note, over 75% of the alerts that indicated that the patient was at risk of developing anaphylaxis or angioedema were overridden. The frequency of opiate-related alerts appears too high. Further research is needed to ascertain how accurate and relevant DAI alerts are for particular patients, and how to present the user with useful drug-allergy alerts.

References:


This study was funded by grant #U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ) Dr. Beeler was supported by the Swiss National Science Foundation
Fine Phenotyping in lung Cancer Using Radiomics and Clinical Data

Luis E. Selva, Ph.D. 1, Saiju Pyarajan, Ph.D. 1,2
1VA Boston Healthcare System, Boston, MA, USA; 2Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA

Abstract
In this abstract we present a novel approach for combining radiomic data extracted from routine clinical images with unstructured and structured clinical data for fine phenotyping.

Introduction
In routine clinical practice radiological imaging modalities like computed tomography (CT) are mainly used either to detect/confirm tumor or to measure changes in tumor size in response to treatment. Recent advances in image processing methods allows for the extraction of numerous quantitative features from clinical images. Aerts et.al (1) recently showed effectiveness of radiomics data as a classifier in both lung and head-and-neck cancer. We used a similar method to couple quantitative features extracted from routine clinical images with structured data derived from pathology and radiology notes using advanced natural language processing (NLP) techniques to attempt to further define clinical phenotypes or sub-phenotypes which could potentially be used as signatures for prognostic prediction.

Conclusion
CT scan images from Non-small cell lung cancer patients were used to extract a number of quantitative features. Structured data were also extracted from Clinical notes. Clustering methods were then used to classify the radiomic and clinical phenotypes. The groups were then compared for differences in tumor progression and response to treatment.

References
1. Aerts et.al Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach NATURE COMMUNICATIONS | 5:4006 | DOI: 10.1038/ncomms5006 |
Cluster Analysis Algorithm For Cohort Comparison

Sunny Shahdadpuri, MBBS, MBA¹, Aditya Sane, MS, BEng¹, Purav Gandhi, MBBS, MBA¹
¹ConvergeHealth by Deloitte, Newton, MA

Abstract
Clinical research on real-world dataset involves a large multitude of clinical data parameters including demographics, diagnosis, surgeries, procedures, prescriptions, laboratory variables as well as side-effects and adverse events. We have developed an advanced visualization to perform a cluster analysis to study relationships between these various data parameters, to create virtually all possible sub-cohort combinations and identify the ones that have significant difference in outcomes.

The Challenge
The explosion of data generated by increasing adoption of electronic medical record (EMR) systems in the last decade enabled by both technology and regulation has created immense opportunities for researchers. Life science researchers are analyzing large volumes of data across various categories like diagnosis, treatments, procedures, prescriptions, and therapeutic pathways. With a large number of data element categories, it is difficult to understand which data attributes drive differences in key outcomes. There is an impending need to be able to rapidly elucidate these drivers and identify the corresponding patient populations, so that medicine delivery could be more precise and targeted.

Method
We have developed a system for analysis and visualization of multitude of clinical variables in context of a clinical study. This system allows for prioritization of variables based on importance, assessment of impact of these variables, and their relationships on end outcomes and quantification of the impact using industry standard comparison tests. There is also a provision for iterating on the results by selecting different cohorts for analysis. The steps followed in our system of analysis are as follows:

- Selection of cohorts and outcomes based on the disease areas and study objectives
- Determination of a top list of attributes ranked according to statistical relevance using the Maximal Information Coefficient (MIC)²
- Cluster formation based on different combination presence or absence of up to twenty clinical variables for the selected cohorts of patients³
- Tracking of each of the clusters for the selected outcome and assessment of statistical difference between the outcomes values for the two cohorts. Statistical difference can be interpreted through a multitude of tests including mean variance, odd ratios, log-odds, or Yule’s Q.

Results
As a result of the above four step process, we have a robust system for cluster analysis to deal with the multitude of clinical variables for each patient, which helps with:

1. Identifying the most relevant clinical parameters for any clinical study
2. Analyze the impact of these parameters as well as their relationships on the outcomes for cohorts of interest
3. Measure statistical significance of the comparison
4. Repeat steps 1 to 3 for a different set of cohorts in a real-time analysis

Discussion
With this concrete approach for comparative statistical analysis based on clustering in mind, we are trying to broaden the applicability of the cluster analysis framework to longitudinal analysis as well as to data sources like claims. We are also in the process of bringing in other statistical tests and modeling techniques as a supplement to the cluster analysis.

A Framework for Assessing Clinical Data Suitability for Observational Study

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

Introduction
With the growing availability of electronic health record (EHR), more and more researchers are starting to use clinical data in EHRs for observational studies. The lack of precise characterization of clinical databases and well-accepted measures of their suitability for observational studies is one of the major barriers to sharing such data for facilitating Big Science. This study aims to develop a framework for characterizing clinical databases and evaluating their suitability for observational studies. It was developed through a literature review that identified common data needs in clinical databases for observational studies by clinical researchers, and frequent designs of clinical databases for research reuse.

Methods
We surveyed the literature on two topics: one on population-based observational studies using clinical databases and the characteristics of these databases that are pertinent to researchers and the other on desiderata (desired characteristics) of clinical databases for understanding the challenges facing research reuse of clinical databases. Fifty-three relevant full-text articles were selected from the 228 retrieved from the search and reviewed for author characterizations of data sources and frequent data limitations, and database design challenges for research reuse. By combining the two review topics derived sub-measure sets, and grouping them into different measures, a conceptual framework was developed to capture the aspects that should be considered for choosing appropriate clinical databases for conducting observational studies.

Results
We identified 21 sub-measures from observational studies utilizing a clinical database, and 35 from desiderata studies. Eighteen of those are shared from both study groupings. A total of 38 sub-measures were grouped into 16 measures that are conceptualized as five categories: policy and administration, relevancy, descriptive metadata and provenance documentation, usable, quality (Table 1).

<table>
<thead>
<tr>
<th>Characteristics for Research Reusable Clinical Databases</th>
<th>Measure (number of finer-grained sub-measures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy and administration: deliberated principles to guide researchers to access clinical data.</td>
<td>Data policy documentation (1), Administrative platform (1), Technical accessibility (1)</td>
</tr>
<tr>
<td>Relevancy: To evaluate the data and pertinent information contained are appropriate for research of interest.</td>
<td>Healthcare organization description (1), Data modeling documentation (2), Research data inventory (1), Retrieveable temporal information (3)</td>
</tr>
<tr>
<td>Descriptive metadata and provenance documentation: documented metadata to identify a resource and describes its intellectual content and information origin</td>
<td>Data provenance (1), Database content synopsis (5)</td>
</tr>
<tr>
<td>Usable: data is able to be used for research of interest</td>
<td>Data representation (3), Usefulness (3), Cohort availability (4), Data linkability (2)</td>
</tr>
<tr>
<td>Quality: an essential characteristic that determines the reliability of data for research.</td>
<td>Data quality control (1), Database data quality (5), Research sample data quality (4)</td>
</tr>
</tbody>
</table>

Conclusion and Discussion
A suitability framework was developed to assess the appropriateness of clinical databases for observational studies. The framework enumerates all the factors that a researcher might want to consider in searching for and selecting a clinical database to use in their research. This framework was built with the perspectives of data custodians and data consumers in mind. For evaluation, we will perform a national survey to validate the framework and elicit experts’ opinions on important metrics. To implement the prototype of the framework, we plan to extend the OHDSI ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems) tool to offer potential database users the ability to interactively review the databases for their research.

Acknowledgments This project was funded by NLM grant 3R01LM006910-15S1 (PI: Hripcsak).
Drug Database Refinement Using Machine Learning and Text Analysis

Reza Sharifi Sedeh, PhD¹, Xianshu Zhu, PhD¹, Yugang Jia, PhD¹, Joey Liu, BSc¹, Oladimeji Farri, MBBS, PhD¹, Daniel Elgort, PhD¹
¹Philips Research North America

Abstract
One of the main challenges in utilizing Healthcare databases is the existing non-uniformity across the databases. We implemented an NLP- and machine learning-based approach to standardizing and mapping drug names to corresponding HICL codes. We applied a Naïve Bayes method to the intensive care unit (ICU) medication database of Philips eRI database, which is pre-processed and cleaned with a natural language processing (NLP) pipeline. Following our evaluation on a sample data set, we have achieved 99.7% of accuracy in predicting correct, uniform drug name entries.

Introduction
Naïve Bayes [1] is a popular and widely used machine learning approach for text classification. However, when there are errors (e.g. misspellings) and redundancies within text data, the performance of machine learning techniques (including Naïve Bayes) can dramatically decline. We hypothesize that machine learning techniques in combination with NLP, as a preprocessing task, would greatly improve the performance of algorithms to accurately classify text for the purpose of refining and standardizing large and 'messy' medication databases for effective down-stream use in clinical and pharmaceutical research.

Methods and Results
We implemented an NLP- and Naïve Bayes-based approach to standardize the Philips eICU Research Institute (eRI) medication database, comprising therapy information over two million patients in over 350 US Hospitals with large amounts of data entry errors and redundancies, and map the drug names to corresponding HICL (Hierarchical Ingredient Code List) codes. Using the Philips proprietary NLP infrastructure which references several ontologies and a home-grown pharmaceutical abbreviation database, we carried out the NLP-based drug name cleaning via four steps: removal of doses and units from the drug names using regular expressions (e.g. 20mmol KCL = KCL), replacing the non-standard abbreviations with full-length drug names (e.g. Fent = Fentanyl), correction of misspellings by referencing a Lucene-based spell-check dictionary of n-gram indexed terms from SNOMED CT, RxNorm, NDDF and general English vocabulary (e.g. Dilauded = Dilaudid = Hydromorphone), standardizing the drug names by including appropriate white spaces (e.g. Calcium gluconate = Calcium gluconate) and removing non-distinguishing keywords (e.g. Propofol volume = Propofol).

The 'cleaned' data resulting from the NLP pipeline becomes the training set for a Naïve Bayesian estimator which uses the corrected (pre-processed) drug-names and their corresponding class indices, HICL sequence numbers, which are a universal and standardized nomenclature for medications and their ingredients. During the training process on a training set of over 68 million drug entries and their corresponding HICL sequence numbers, two important parameters are calculated: (i) the percentage frequency of a drug name appearing within a given HICL class and (ii) the percentage frequency of a specific class (HICL sequence number) appearing in the whole sample. These two parameters are used later during the testing phase, where the probability of a given drug name belonging to a HICL class is calculated. As a result, given the drug name, the class that has the highest posterior probability is selected. One of the authors (OF) with clinical and pharmacological expertise manually reviewed the performance of our approach on the most frequent entries (~93%) of a table of over 57 million drug entries, which is comprised of 89.5% drug name entries and 10.5% non-drug name entries. Our approach is 99.7% accurate in predicting correct, standardized drug name entries (Table 1). It is also noteworthy that our approach was only targeted at standardizing the drug name entries, and not the non-drug name entries such as Normal Saline and Packed Red Blood Cells.

Table 1. Manual evaluation of our proposed approach on a sample data set.

<table>
<thead>
<tr>
<th>Type of entry</th>
<th>% of the sample data set</th>
<th>Correct (%)</th>
<th>Wrong (%)</th>
<th>Not recognized (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>89.5</td>
<td>99.7</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Non-Drug</td>
<td>10.5</td>
<td>4.5</td>
<td>1.3</td>
<td>94.2</td>
</tr>
</tbody>
</table>

Reference
1. Lewis DD. Naïve Bayes at forty: the independence assumption in the information retrieval. ECML 1998; Chemnitz, German.
Registries in Accountable Care: Essential Data Management in New Models of Care

John Sharp, MSSA\textsuperscript{1,2}

\textsuperscript{1}HIMSS, Cleveland, Ohio; \textsuperscript{2}Kent State University, Kent, Ohio

Abstract

Registries are a powerful informatics tool for research and public health. As both commercial payers and the Centers for Medicare and Medicaid Services work to shift incentives toward value-based purchasing, demand for reliable, accessible data on populations is growing. The purpose of this poster is to define accountable care organizations (ACOs), explain the importance of registries in managing data for ACOs, and discuss specific informatics requirements unique to accountable care registries.

Introduction

Registries have traditionally been used for research to study conditions and treatment groups, public health to monitor disease and epidemics and quality registries for public reporting.\textsuperscript{1} Registries for accountable care have some similarities to these registries but a different focus and incentives: to manage a population based on an insured cohort and incentivized by risk-sharing insurance plans. Quality reporting is a requirement for risk-based contracts as such, quality measures are key data points.\textsuperscript{2} These registries also require risk-stratification methods to facilitate care coordination.

Designing ACO Registries

As with any registry, the purpose and objectives must be designed at the onset. Patient cohorts are typically defined by the risk contract from Medicare or insurer. Data elements need to include data that enables risk-stratification and quality reporting.\textsuperscript{1} Sources of data include electronic medical records, insurance claims and patient questionnaires. Data architecture must be flexible as needs within the new care model can change quickly. Data integration from these sources continues to be a challenge especially for patient matching, data migration and entity disambiguation.

Transforming Registry Data in Knowledge

To make the data actionable for clinical decision support and care coordination, dashboards and data visualization tools are essential. Self-service analytics for all members of the care team assist in identifying new risk factors and subgroups that require intervention. The next level of transforming this data is predictive modeling; this is especially important in predicting risks in chronically ill populations, which may result in unnecessary emergency room visits or hospitalizations. Informatics tools, such as, integrated decision support based on these models, can improve care.

Future Role for Use of Registries in Accountable Care

In the dynamic nature of accountable care, structure and data needs are constantly changing. New approaches to obtain and analyze data near real-time are challenges for these registries. Quality metrics can vary from one payer contract to the next. Standardization of these metrics will enable more efficient use of registries in the future. Currently a centralized repository of best practices is lacking and would be helpful as ACOs expand nationally. Innovation should be encouraged to solve these and other issues as they arise.

Conclusion

Registries for ACOs are a key informatics tool for the success of this care model. While there are similarities to other types of registries, accountable care registries have unique requirements including near real-time data, specific quality measures, support for care coordination and risk stratification.

Support by AHRQ contract #HHSA290201400004C

References

MedBus: A service-oriented architecture for enabling the research data economy

Amitava Shee, BE¹, James Estill, PhD¹, John Brussolo, MBA¹
¹University of Michigan Medical School Information Services, Ann Arbor, MI, USA

Abstract

Researchers often require access to multiple data repositories to carry out their work. Although these data sources are a valuable asset to the community, researchers are often left to find these resources and navigate access to these data on an individual, point-to-point basis. In order to accelerate data discovery and access we developed a service-oriented architecture platform, MedBus, for securely self-provisioning these resources. MedBus will ultimately enable federated access to over 300 enterprise data repositories.

Problem

Modern research programs often require integration of high velocity data from heterogeneous sources within a variety of analytical platforms. Researchers at our institution utilize data from over 300 enterprise repositories and are required to integrate these resources on an individual point-to-point basis. Furthermore, discovery of these data repositories can be difficult, and the process of how to negotiate access to these resources can be opaque. Although centralized data warehousing in platforms such as i2b2 can facilitate data access and interoperability, there also exists a need for data federation systems that can respond to the dynamic requirements of big data and enable use cases that require integration independent of any single platform.

Solution

To address the needs for a secure, federated solution supporting research, we developed MedBus (http://goo.gl/1Bg4rc), a service-oriented architecture (SOA) that addresses the current and emerging challenges of research data discovery, access, and integration. The MedBus enterprise service bus presents research data as a collection of interoperable data services that are agnostic to the technology in the underlying data repositories. These services facilitate user-oriented mashups of research data and provide an extensible platform for increasing the visibility and utility of enterprise data resources. Since MedBus has been built from open source components, individual data owners can develop services that provision data within the MedBus framework, and the broader research community can be engaged in the extension of the MedBus platform.

Results and Conclusions

The initial deployment of MedBus includes services that provision data from: 1) a physiological waveform metadata store (5GB) and 2) the laboratory information management system (LIMS) used by our institutional biorepository (currently 148K specimens from 25K subjects). These services are consumed by a locally developed cohort discovery and data delivery tool that provides a secure and intuitive interface for complex queries that connect data from these services with patient data from our institutional EHR warehouse. This has resulted in accelerated access to research data in a self-service environment that facilitates the transition from structured data to actionable information. Current MedBus development includes the addition of a data access attestation service and eConsent that integrates with IRB approval workflows. We are also incorporating data virtualization as a method to accelerate development of new data services.
Data Mining to Predict Healthcare Utilization in Managed Care Patients
Lincoln Sheets, MD1,2, Michael A. Phinney, BS3, Sean A. Lander, MS1,
Jerry C. Parker, PhD2, Chi-Ren Shyu, PhD1,3,4
1University of Missouri, MU Informatics Institute, Columbia, Missouri
2University of Missouri, School of Medicine, Columbia, Missouri
3University of Missouri, Computer Science Dept., Columbia, Missouri
4University of Missouri, Electrical and Computer Engineering Dept., Columbia, Missouri

Problem Addressed. Predictive analytics can be used to rapidly spot hard-to-identify opportunities to better manage care1, but most current risk prediction models fail to reach 80% accuracy2, but data mining algorithms would permit models that surpass the size limits of traditional statistical methods3, and could use the rich data available in the EHR4.

Methodology. From a sample of 10,000 primary care patients, those with the highest charges during one year were compared to an equal number of patients with the lowest non-zero charges. Clinical attributes were associated using the FP Growth algorithm, and validated by multiple regression.

Results. Of 19,012 EHR data elements, 9 predicted charges over $75,000 (p<0.20).

| EHR Code | Description | Pr > |t| |
|----------|-------------|------|---|
| 786.6    | Swelling, mass, or lump in chest | 0.0368 |
| V72.84   | Pre-operative examination, unspecified | <0.0001 |

| --- ICD-9 Diagnostic Codes --- |

| --- Diagnosis-Related Groups (DRG) --- |

| 358 | Digestive system procedures without complicating conditions | 0.0524 |
| 571 | Skin debridement with complicating conditions | 0.0242 |
| 640 | Disorders of fluids or electrolytes with major complicating conditions | 0.0015 |

| --- SNOMED Codes --- |

| 493070016 | Ischemic heart disease | 0.0052 |

| --- Therapeutic classes of prescriptions --- |

| 5-HT3 antagonists | <0.0001 |
| CD20 monoclonal antibodies | 0.1026 |

| --- Other clinical attributes --- |

| Body Mass Index (BMI) < 18.5 | 0.0044 |

Conclusions. Systematic mining of clinical attributes from the electronic health records of adult primary care patients reveals multiple predictors of high healthcare utilization. Association mining narrows feature sets that are otherwise too large for linear regression, with good resulting overall significance.

References.

This publication was supported by Grant Number 1C1CMS331001-01-00 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services or any of its agencies.
A Continuous Markov Model Approach Using Individual Patient Data to Estimate Mean Sojourn Time of Lung Cancer

Shiwen Shen\textsuperscript{1,2}, Simon X. Han\textsuperscript{1,2}, Panayiotis Petousis\textsuperscript{1,2}, Frank Meng\textsuperscript{2} PhD, William Hsu\textsuperscript{2} PhD, Alex A.T. Bui\textsuperscript{2} PhD

\textsuperscript{1}Department of Bioengineering, \textsuperscript{2}Medical Imaging Informatics (MII) Group, Department of Radiological Sciences; University of California, Los Angeles, CA, USA

\textbf{Introduction.} Large amounts of existing periodic screening data for chronic diseases offer the potential in to better understand disease dynamics and progression. This knowledge is able to provide individualized recommendations for screening and diagnostic tests. One major challenge of leveraging this data is that there are large amounts of missing or partial observations. For instance, patients found to have lung cancer might receive treatment/intervention, thereby obviating further observations on lung cancer progression. To overcome this problem, we present an approach based on the continuous-time Markov model (CMM) and apply it to estimate lung cancer progression, specifically the average duration time in the preclinical screening detectable state before progressing to the clinical symptomatic state. This duration is also known as the \textit{mean sojourn time} (MST) and provides an absolute upper limit to the obtainable lead time. MSTs can vary significantly across screening approaches, reflecting the efficacy of the screening process. Low-dose computed tomography (LDCT) and chest x-ray (CXR) are two important modalities for lung cancer screening. The National Lung Screening Trial (NLST) was a randomized multicenter study that compared these two imaging modalities’ ability to detect lung cancer in an at-risk population (older current and former heavy smokers). In this work, we use the NLST dataset to estimate MSTs for LDCT and CXR. Our approach assumes the individual resides in one of the predefined states and can capture continuous time information between states. This work describes the first time that a CMM-based approach is applied to estimating MSTs for lung cancer under LDCT and CXR using individual-specific transition time information for a large population dataset.

\textbf{Methods.} We develop a three-state homogeneous continuous Markov model for the LDCT screening arm. For LDCT, the three states are respectively defined as: State 1, no detectable cancer; State 2, preclinical cancer; and State 3, clinical cancer. Individuals can only transition from State 1 to State 2, and from State 2 to State 3 (as shown in Figure 1). The transition rate from the non-detectable state to the preclinical state is $\lambda_{1,2,LDCT}$, and similarly the transition rate from the preclinical to the clinical state is $\lambda_{2,3,LDCT}$, where $1/\lambda_{2,3,LDCT}$ is the MST. Based on the transition rates a matrix is formulated comprising finite transition probabilities $P(t)$. Each element of this matrix represents the transition probability of an individual case across the states of the Markov model. NLST participants underwent three rounds of screening. For each round of the screening, screen-detected lung cancers are assumed to be at the preclinical state, and negative cases are assumed to be at State 1. Interval cancer cases are assumed to be at the clinical state. The transition time between states for each individual is accounted for in this model (24,988 subjects in CXR arm and 25,473 subjects in CT arm). Post-screening reported cancers are not included in this study. We initially assume that the sensitivity of each screening is 100%. Maximum likelihood estimation (MLE) is used to estimate the parameters and solved using a quasi-Newtonian optimization method. The MST for specific patient cohorts is estimated by using a proportional hazards model to relate the transition intensities to covariates (e.g., gender). A similar approach was used to develop the model for the CXR arm.

\textbf{Results.} Our estimated MSTs are: 0.7 years (95% confidence interval: 0.61-0.83) for CXR and 1.76 years (1.58-2) for LDCT. For LDCT, the female and male MSTs are 1.94 years (1.10-3.69) and 1.64 years (1.10-2.62); the CXR MSTs are 0.67 years (0.41-1.15) and 0.82 years (0.58-1.22). Our estimations are comparable to reported MSTs for lung cancer in recent studies (1.33-6.08 years for CT and 0.45-2.24 years for CXR).

\textbf{Conclusion.} In this work we estimate a longer MST for patients who underwent LDCT screening compared to CXR in individuals at high risk, suggesting that LDCT is more effective in detecting early stage lung cancer. The CXR MST is less than 1 year, suggesting a smaller window of opportunity for lung cancer screening. There is no significant difference between MSTs of females and males in both LDCT and CXR.

![Figure 1. Illustration of the three-state continuous Markov model for low-dose computed tomography.](image-url)
Identifying population characteristics tables in full text articles
Garrick Sherman, MS, Catherine Blake, PhD, Jooho Lee, MS
University of Illinois, Urbana Champaign, IL, USA

Abstract: Authors frequently use tables to provide population characteristics of patients who participate in a study. Automated indexing methods tend to focus on the narrative in an article, rather than tables, even though this data can help physicians determine whether a study is relevant to their current clinical encounters. We provide an automated method to identify tables that contain population characteristics, which can be extracted in a later step.

Introduction
Several efforts have focused on standardizing the data reported in a clinical trial (CONSORT), a meta-analysis of observational studies (MOOSE), and an epidemiological study (STROBE). These new guidelines do not address how to extract information from the millions of already published studies. Our goal is to identify and extract demographics (e.g. age, gender, and ethnicity), behavioral factors (e.g. tobacco and alcohol consumption), and medical descriptors (e.g. disease stage) from tables, such as that shown in Figure 1. As a first step, we identify tables containing data corresponding to the patient aspect in PICO [1] and population, intervention or risk factor, and medical condition of the information synthesis framework [2].

Materials and Methods
Full text articles in 18 breast cancer journals from the open access subset of PubMed Central were downloaded. Tables were extracted from the 3,638 NXML files; only the first table is included in these experiments. A heuristic approach to processing tables, similar to [3], was used. Headers were identified using HTML tags, and factors were defined as the first non-empty cell in a row that included either a) only one non-empty cell, or b) at most half as many non-empty cells as total cells in the row (including empty cells). Once factors were identified, the first cell of the next row was interpreted as a level and all subsequent cells in that row were considered values. Differently structured tables were labeled with a generic “Main” factor, but otherwise parsed identically.

The task of identifying population characteristic tables was framed as a classification problem. Several features were considered: factor unigrams, factor and level unigrams, caption unigrams, and numbers. Only unigrams were considered due to the brevity of table text. Text was tokenized on whitespace and converted to lower case (no stemming was applied since terms were not expected to exhibit many forms). Stopwords were eliminated, as were terms occurring in ≤3 or >95% of the tables. A training set of 1,001 tables was annotated by hand (414 population tables), and information gain (see [4]) was used to identify the most discriminating terms. Four classifiers—support vector machine (SVM), naïve Bayes (NB), decision tree (DT), and the general linear model (GLM)—were used to classify the test set in the Oracle Data Miner (with default settings). The features identified in the training set were then used on an unseen test set comprising 497 tables.

Results
NB outperformed the other classifiers for all feature combinations, with a maximum accuracy of 87.6% on the judged subset, which included 50 random positive and 50 random negative examples. The most effective set of features were the top 50 terms, no captions or numbers and two columns (i.e. using terms from both the factor and levels). The NB model was applied to the test set. There were 29 false positives and 32 false negatives leading to an overall accuracy of 87.73% in the test set.

References
Performance Comparison of Running Clinical Rules in Drools and Plain Java Implementation

Jian Shi, MD, MS, Erik Smith, Thomas J. Van Gilder, MD, JD, MPH
Transcend Insights, San Diego, CA

Introduction

Drools is an open source, Apache Foundation, community-based project that provides an integration platform for the development of knowledge-based systems. It is developed in Java and has a modular architecture which is based on an object oriented implementation of the PHREAK – a lazy rule matching algorithm to enable Drools to handle a larger number of rules and facts. Drools has the following major advantages: Declarative Programming, Logic and Data Separation, Speed and Scalability, Centralization of Knowledge and Understandable Rules. It can also be easily integrated with other open source frameworks such as Spring and Apache Camel, among others. Because of those advantages, Drools has been used in healthcare systems to process clinical rules. It is the core component of OpenCDS which has numerous collaborators like Intermountain Healthcare, Wolters Kluwer Health and others. Compared to rules in other industries, clinical rules tend to have more complicated logic, especially to implement healthcare quality measures. No study has been found regarding the performance of Drools to process clinical rules.

Methods

Drools version 6 was installed and two other services were implemented to feed the patient data and value set codes to the Drools rule engine. The childhood immunization measure (10 different vaccines) from HEDIS was then written in Drools. The eligibility was written in one rule, and the compliance and optional exclusion for each vaccine were written in separate rules. The same rule logic was also hard coded directly in Java. The rules were grouped by each vaccine. The run time was then measured against the number of vaccines that were analyzed on each run. The plain Java implementation shares the same architecture except that the engine component is implemented in plain Java, and reuses the patient data and value set services. Both implementations were run against the same 10,000 patient claims database with about 300,000 records. The output was then compared.

Results

There is a learning curve to implement Drools and writing rules in Drools rule language (drl). The age requirement can be easily done while continuous coverage, value set matching and counting unique service dates that are slightly challenging due to the limitation of the operator/functions packaged in Drools. This requires Java coding either in utility class or the rule “then” section. When comparing the results, two different implementations generate the same output. The data loading which was implemented in JDBC takes a significant portion of overall run time. The comparison of performance was done with the exclusion of data loading time. The results are shown below.

Conclusion

Simple requirements like age range or a look back window can be easily implemented in the drl file. Any update to those requirements can be done within the rule grammar without making any changes to the engine coding. The rule execution time on clinical rules with certain complexity is impressive, yet it is still longer than the plain Java implementation. For a use case in which continuous updates occur, Drools is a good option to consider. However, if performance (as measured by execution time) is the main metric, it may be worth researching other implementation options.
Implementation of a Mobile Electronic Medical Record System that Uses a Problem-oriented Contiguous Timeline View

Akiko Shibuya, RN, PHN, PhD1, Keisuke Ogawa, MS2, Kazunori Matsumoto, MS, PhD2, Masayuki Hashimoto, MS, PhD2, Yukihiro Maeda1, Yoshiaki Kondo, MD, PhD1

1Department of Health Care Services Management, Nihon University School of Medicine, Tokyo, Japan; 2KDDI R&D Labs, Saitama, Japan

Abstract
A mobile electronic medical record (mEMR) system was designed that uses a Problem-oriented Contiguous Timeline View (PCTLV), which presents an overview of patient information (e.g., prescription drugs, laboratory results) along a time-axis on a single screen. Clinicians can view medical history (e.g., insulin doses, glucose levels) that relates to disease management while easily changing time-scales (day, month, year) with one click. The mEMR that uses a PCTLV may have a significant impact on clinicians’ decision-making.

Introduction
A clinician’s work is information- and communication-intensive and highly mobile. The potential for a mobile electronic medical records (mEMR) system, accessible on mobile devices such as smartphones and tablets, to improve patients’ outcomes is widely recognized. However, concerns still exist regarding barriers to mEMR use by clinicians. In a previous study, poor information overview (e.g., prescription drugs, laboratory results) and poor usability for the decision-making process were considered important barriers for clinicians1. To help address these issues, we developed an mEMR that uses a Problem Contiguous Timeline View (PCTLV), which enables visualization of patient information (e.g., prescription drugs, laboratory results, disease diagnosis) along a time axis on a single screen.

Methods and results
Figure 1 shows an overview of our mEMR system that uses a PCTLV. With this system, clinicians can view patient information (e.g., prescription drugs, laboratory results) while easily changing time-scales (e.g., hours to days, days to months) with one click on a single screen (Figure 2). Furthermore, when a clinician selects a term of interest (e.g., prescriptions, test results, clinician notes), a new window appears showing detailed information (e.g., dosage, laboratory data, and computed tomography).

Figure 1. Overview of the system

To verify the efficacy of the system, 14 clinicians within a hospital (approximately 1,300 beds, 50 clinical units, and 2,300 employees) in Japan participated in this trial. They viewed simulated patient data and their real patient data on our mEMR during the pilot test period. After this procedure, we assessed system usability using a questionnaire rating items on a five-point scale and a semi-structured interview. The information overview provided by this new mEMR was significantly higher than that of the existing EMR system (mean, 4.0 vs. 3.0 point, p=0.05). Clinicians believed this system was important in decision-making, as it was useful that all information on their patients’ drug and examination history could be obtained on a single screen. This indicates that an mEMR that uses a PCTLV was positively received.

Conclusion
The present study suggests that an mEMR that uses a PCTLV may have an important role in supporting better clinical decision-making.

References
Improving Failure Mode and Effects Analysis through Electronic Health Record-Assisted Team Identification

Gayle Shier Kricke, MSW¹, Matthew Carson, PhD¹, Young Ji Lee, PhD¹, and Nicholas Soulakis, PhD¹
¹Northwestern University Division of Health and Biomedical Informatics, Chicago, IL

Introduction
Failure Mode and Effects Analysis (FMEA) is a proactive risk assessment approach often used to inform healthcare quality and safety improvement. FMEA utilizes a team-based approach with expert facilitation to analyze the different ways a process may fail and to prioritize interventions addressing identified possible failures (failure modes). FMEAs are popular in healthcare because they satisfy The Joint Commission’s hospital accreditation Standard LD.04.04.05, EP 10, which requires a proactive risk assessment of a high-risk process at least every 18 months. Identifying failure modes and effects draws upon team members’ experience rather than data alone, which makes the outcome of the process largely dependent on the people at the table. Inaccuracies and inconsistencies occur if team representation is not appropriate or comprehensive.

Objective
The objective of this study was to determine whether electronic health record (EHR) documentation could be leveraged to develop a comprehensive list of providers to be represented in an FMEA process. We hypothesize that a query of EHR documentation will identify providers highly involved in healthcare processes who would be overlooked when assembling a FMEA team through typical, non-EHR driven methods.

Methods
Data Source and Variables
Data was derived from the Northwestern Medicine Enterprise Data Warehouse (NM EDW) for 3,120 Emergency Department (ED) encounters at Northwestern Memorial Hospital (NMH) between November 1, 2013 and October 31, 2014. A process map of ED workflows was created with input from ED staff. Each step in the process was mapped to documentation in the NM EDW. Variables used for analysis included provider name, provider type, activity type, and number of times the provider completed the activity during the sample of encounters.

Analysis
NM EDW documentation was then used to identify providers involved in ED workflow steps, which was compared to the process map-derived list of team members. The EHR-derived list of team members was examined to determine how the information could be used for FMEA team construction.

Results
EHR data identified 23 different process steps performed by 2,617 individual providers across 21 different provider types for the sample of ED encounters. Twelve additional provider types not included on the hand-drawn process map were identified, including two types that performed more than 5 process steps, which suggests more people may be involved in ED care than typical team assembly would identify. Within each activity, it was possible to see which individual providers performed each process step most often, which could assist in selecting the most informed providers to be invited to the FMEA process. When providers performed steps, they did so an average of 15.6 times, but individual providers with experience far exceeding the average were identified.

Conclusions
EHRs may be a useful tool in supplementing current FMEA team assembly practices by identifying engaged providers overlooked by typical process mapping, which could help improve FMEAs’ accuracy and applicability by informing the development of a more comprehensive FMEA team. Additionally, EHRs may help in prioritizing individuals to be invited to the FMEA based on their experience with process steps. Hospital quality professionals may benefit from tools that make information from the EHR quickly available for improvement planning and decision-making.
An Enterprise Clinical Data Pipeline for a Cancer Center
Emily Silgard, MS1, Tony Galuhn, BA2, Kathryn Egan, MS1,
Adam Rauch, BS2, Paul Fearn, MBA1

1Fred Hutchinson Cancer Research Center (Fred Hutch), 2LabKey Software, Seattle, WA

Introduction
To implement a scalable system for acquiring and processing unstructured or semi-structured clinical data for 350,000 historical patients and over 6,000 new cancer patients treated within the Fred Hutch/University of Washington Cancer Consortium annually, we designed an enterprise clinical data pipeline. This pipeline is meant for deployment of natural language processing (NLP) algorithms to extract data elements from clinical narratives, and for assignment, tracking, and completion of manual data abstraction, quality assurance, and training data creation.

Previous Work
This project provides a framework for developing and deploying any kind of NLP algorithm - commercial, open source, or homegrown - and for integrating annotation tasks, facilitating manual data abstraction and developing training data for NLP algorithms. Existing NLP platforms for processing unstructured text (e.g. cTAKES1) are very extensible but less precise within the cancer domain and do not include parallel manual annotation and quality assurance task management. Annotation toolkits (e.g. Brat2 and Knowtator3) do not have built-in NLP algorithms or tools. The GATE4 platform can be used for automated processing and annotation, but while it is a useful development tool, it would not be as effective for large-scale, ongoing data entry for a variety of users. Commercial NLP packages (e.g. Linguamatics5) offer many out-of-the-box options, but limit overall freedom of data models, algorithms, and application development and do not provide a platform for the creation of quality labeled data.

Methods & Materials
We have developed the data acquisition pipelines and the front-end applications that provide data exploration and visualization portals for the Hutch Integrated Data Repository and Archive (HIDRA) using LabKey Server6, an open source platform that helps translational research teams integrate, analyze, and share clinical data. We use LabKey Server for instantiating the NLP engine and parametrizing the processing pipeline. The pipeline has a hierarchical design where each level is linked to a separate directory containing python scripts/modules, text file gazetteers, dictionaries and metadata needed to extract the relevant set of data elements. The first tier corresponds to document types and the second to disease groups. Document types can differ greatly in discourse and data elements. They may come from disparate clinical source systems, and may have varying formats and metadata, requiring customized parsing and pre-processing. The current design will allow us to take advantage of existing NLP tools and algorithms where relevant, while having the flexibility to develop our own algorithms for enterprise specific definitions and data elements which do not have existing, well-established NLP algorithms.

Testing and Evaluation
We will evaluate each function of the pipeline with end users (developers, abstractors, researchers) through common workflow scenarios. Performance of the NLP algorithms will be evaluated for the pilot against data entered by trained medical abstractors for a subset of records for which results and detailed error analysis will be reported.

Discussion
Next steps include extending the pipeline to other cancer types, ideally involving minimal development of new, disease-specific algorithms and dictionaries, as well as extension to other document types. The enterprise pipeline would ideally be portable to any group or organization utilizing a similar information model and any existing or home grown NLP algorithms could be implemented within its framework. While initial parsing and preprocessing of clinical documents would have to be tailored to match organization-specific sources and formats, the bulk of the data elements needed for cancer research remain relatively similar throughout the cancer research community.

References
Abstract: Identification and subsequent addition of elements to a “home-grown” ambulatory electronic health record (EHR) system can yield useful resources for primary care providers in Patient Centered Medical Homes (PCMH) in the areas of care planning, patient self care support, and new medication management. Opportunities for the EHR to support PCMH goals were identified with new system functionalities added to help practices bridge gaps in documentation needed to achieve National Committee for Quality Assurance (NCQA) Recognition.

Introduction: Primary care practices seeking recognition as Patient Centered Medical Homes must undergo a rigorous process and meet thresholds for particular elements, some of which are assessed through chart review. In this study, we identified areas in which practices seeking PCMH recognition had difficulties meeting certain documentation requirements. In response, changes were made to the EHR to better support these functions. We analyzed the transformation of practices once these new functionalities had been implemented to gain insight to the benefits of these new features.

Methods: By conducting pilot chart reviews of eight patients per practice, we identified four areas of difficulty meeting documentation requirements for 22 primary care practices during 2014 in preparation for a subsequent full chart review as part of the NCQA application. These included developing care plans with treatment goals, providing patients with self management plans with goals, and providing information about and assessing patient understanding of new medications. We worked with the practices as well as with HIT developers to make modifications within the EHR system, which included templates to satisfy care plans and self management plans with goals, as well as “Patient Educated” and “Patient Understands” checkboxes for prescribed medications. Practices were encouraged to use the new functionalities for documentation to meet these factors. We then conducted full chart reviews of 48 patient charts per practice. We assessed and compared the pilot and full chart review findings for the practices.

Results:

Table 1. Changes in percent of charts passing documentation requirements pre and post new EHR functionality

<table>
<thead>
<tr>
<th>Documentation Requirement</th>
<th>Pre-EHR Intervention</th>
<th>Post-EHR Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care plan with treatment goals</td>
<td>18%</td>
<td>95%</td>
</tr>
<tr>
<td>Self management plan with goals provided to patient</td>
<td>9%</td>
<td>100%</td>
</tr>
<tr>
<td>Information about new medication given to patient</td>
<td>36%</td>
<td>54%</td>
</tr>
<tr>
<td>Understanding of new medication assessed</td>
<td>77%</td>
<td>95%</td>
</tr>
</tbody>
</table>

As shown in Table 1, the percentage of the practices meeting passing criteria for each of the four factors increased after new EHR functionality was implemented.

Conclusion: Modifications to the EHR system tailored to primary care practices can assist in supporting care processes and may ease the burden of documentation requirements to become NCQA recognized as Patient Centered Medical Homes. After EHR system modifications, practices’ scores improved in four important care process areas, making it evident that additional EHR documentation support was a major contributor to improved scores.

References
A New, Touch-screen Sensitive Display for Management of Diuretic Therapy of Heart Failure Patients in Critical Care Setting

Jiri Sklenar, Ph.D., Jay Parikh, M.D., James O. Mudd, M.D.
Knight Cardiovascular Institute, Oregon Health & Science University (OHSU), Portland, Oregon

Abstract

The user experience of providers taking care of heart failure patients can be improved by augmenting existing EMRs with touch screen displays and custom-developed applications. Putting providers in the driver’s seat from the very beginning of the developmental process is essential for the overall success. Our methodology can be applied to other medical areas as well.

Introduction

Providers using current EMRs are confronted with large amounts of data, but paradoxically an intuitive access to clinical information is not always readily available. Deriving actionable information from a multitude of discrete data elements remains a time-consuming, tedious, and potentially error-prone process, due to requisite navigation through a number of screens and menus. This problem becomes compounded when providers are caring for complex multiple co-morbid illnesses such as heart failure patients.

Hypothesis

We therefore hypothesized that using touch-screen technology with standard data interfaces from the EMR, and custom-developed software, we can design an intuitive and responsive tool for busy providers, to assist them in answering important clinical questions almost immediately, without any need to navigate through multiple screens of text and graphs. This was specifically piloted in hospitalized heart failure patients to rapidly assess changes in volume status, weight and renal function in response to a given dose, class and route of diuretics.

Results

Using our application (Figure 1) we were able to more quickly answer clinical questions in seconds, compared to minutes when only tools available in the EMR were used.

Conclusion

Using a portable tablet, we designed a flexible software system allowing for rapid assimilation of multiple data elements into a simple user interface to augment clinical decision-making. The lessons learned are not specific to heart failure patients only, and will be easily applicable to other areas of healthcare in caring for patients with complex chronic medical conditions.

Figure 1. Display for monitoring diuresis includes weight, creatinine levels, and intravenous administrations of diuretics.
Stacey L. Slager¹, MS, Charlene Weir¹², BS, PhD, RN, Heejun Kim³, Javed Mostafa³, PhD, Guilherme Del Fiol¹, MD, PhD
¹Biomedical Informatics, University of Utah, Salt Lake City UT; ²George E. Whalen Veterans Affairs Medical Center, Salt Lake City, UT; ³University of North Carolina at Chapel Hill, NC

Introduction

The purpose of clinical research is to produce evidence for clinicians to use to support patient care. However, using information from clinical studies for clinical decision making requires significant cognitive effort and time. In this pilot study we designed alternate information displays for presenting the results of clinical studies to clinicians. We hypothesized that, compared to traditional narrative formats, a structured display improves clinicians’ ability to scan, judge the relevancy, and interpret the findings of clinical studies.

Methods

We compared two alternatives for displaying clinical studies with standard display provided in PubMed’s search results. The first display provides a narrative summary with the abstract title, journal, publication date, and study conclusions. The second display presents key study elements in a structured tabular format according to the PICO framework (population, intervention, comparison, and outcome)¹. The displays were iteratively designed and presented to members of our research team and target users for feedback. The overall design principle was to present only the information necessary to help clinicians judge the relevancy of a study to a specific patient and to interpret the gist of the study findings. Twelve physicians were invited by email to participate. Participants viewed three brief vignettes along with five studies that were randomly presented in one of the three display formats. The five studies were pre-searched by the study team so that all participants were exposed to the same studies. Participants were then asked to rate each display using an 11-item Likert scale questionnaire (1=strongly disagree; 7=strongly agree), which included items from the Simple Usability Scale (SUS)² and measurements of participants’ ability to judge relevancy and interpret study findings.

Results

All twelve invited physicians participated in the study. The PICO table rated 1.5 to 3.0 points higher than PubMed as the baseline and 0.1 to 0.9 higher than the narrative summary. Averaging all 11 questionnaire items, subjects favored the PICO table compared to the text summary and PubMed (5.9 vs. 5.6 vs. 4.0). Figure 1 shows a subset of the individual rating criteria.

Discussion

A structured format according to PICO elements is a promising way of improving clinicians’ use of clinical studies to support patient care decision making. Limitations include a relatively small sample size, although sufficient given the exploratory nature of the study. A study with a larger sample size is underway.

Acknowledgement

This project was supported by grant 1R01LM011416-01 from the National Library of Medicine.

References

A Novel Anatomical Semantic Ontology for Identification of Anatomically Proximate CTs Using LOINC Codes

Benjamin H Slovis, MD1, Tina Lowry, MS1, Bradley N Delman, MD2, Anton Oscar Beitia, MD3
Gilad Kuperman, MD, PhD4, Charles DiMaggio PhD5, Jason S Shapiro, MD1

1The Department of Emergency Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
2The Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
3The Department of Radiology, Wyckoff Heights Medical Center, Brooklyn, NY, USA
4The Department of Biomedical Informatics, New York Presbyterian Hospital, New York, NY, USA
5The Department of Anesthesiology, New York Presbyterian Hospital, New York, NY, USA

Introduction:
When patients seek care across multiple clinical settings there is a risk of the fragmentation of clinical data, including medical imaging records. Health information exchanges (HIE) are expected to reduce potentially avoidable duplicate imaging, improving the quality and cost of health care. In the future, it may be possible to alert ordering clinicians across an HIE when a potentially avoidable duplicate study is ordered, but more work is needed before alerting rules can be developed based on the anatomical location of studies from multiple sites. When patients receive computed tomography (CT) scans at different institutions they can be categorized as same (the exact same CT based on identical Logical Observation Identifiers Names and Codes (LOINC)), similar (same anatomical region but different CTs/LOINC codes, such as CT head and CT orbits), or proximate (different CTs/LOINC codes but in proximate anatomical regions such as CT head and CT neck), and it may be desirable to identify these as potential duplicates. However, no anatomical framework yet exists for the identification of similar or proximate CTs. We propose a novel anatomical semantic ontology that categorizes LOINC codes for CTs into general anatomical regions, allowing identification of anatomically similar and proximate CTs to facilitate alerting via a HIE-based duplicate detection system, and future analyses employing this framework.

Methods:
The New York Clinical Information Exchange (NYCLIX) was the source for de-identified reports of all patients who received CTs from the six participating sites in the HIE between 3/09 and 7/12. Site-specific proprietary names for each class of CT scan were mapped to LOINC codes for comparison across sites using a previously described mapping process. (Beitia, AMIA 2013) Each LOINC code was then assigned by two of the authors up to five of nine possible independent anatomical identifiers from our novel ontology, based on the type of CT and its anatomical location. These included: head, neck, chest, abdomen, pelvis, left and right upper as well as left and right lower extremities. For instance, LOINC code 44115-4 (Abdomen and Pelvis CT) was assigned two anatomical identifiers for abdomen and pelvis, as was 24964-9 (Spine Lumbar CT W contrast IV). Inter-rater reliability was analyzed using the Kappa statistic. Twelve additional common combinations of proximate anatomical identifiers were recognized, such as “chest and abdomen”. This system resulted a total of 21 possible individual or combinations of anatomic identifiers. Those studies that lacked appropriate description were labeled unknown and excluded. For the extremities, when no laterality was specified, the study was presumed to be bilateral. The number of successful matches was then calculated for the individual or combined anatomic identifiers.

Results:
717,231 CT scans were mapped to 217 different LOINC codes. Twenty-three (10.6%) of these LOINC codes, accounting for 8,252 (1.15%) CTs were excluded because they lacked appropriate description and the anatomical region was unknown. This resulted in 708,979 CTs mapped to 194 LOINC codes in our dataset that underwent anatomic identifier assignment. Inter-rater reliability was found to be $\kappa=0.89$ (p<0.001). The most frequently imaged region of the 9 independent anatomical identifiers was the head, which accounted for 262,121 (27.3%) of all studies. When comparing the 21 independent or combination anatomical identifiers, the head was included in 248,741 (35.08%) of studies while combined abdomen and pelvis accounted for 174,966 (24.68%) of studies.

Conclusions:
Our novel anatomical semantic ontology should aid in the identification of similar and proximate CT scans across multiple sites in an HIE, laying the groundwork for future HIE-based alerting systems to inform clinicians that same, similar or proximate prior CTs exist for a patient. Such a system could potentially reduce avoidable duplicate imaging, and improve the quality and safety of care while reducing costs. Next steps include an analysis of the actual rates of same, similar and proximate CTs across multiple sites in a large HIE dataset.
Leveraging Genetic Findings to Identify High-Risk Chronic Kidney Disease in the Electronic Medical Record

Craig Smail, MA MSc
1Biomedical Informatics Training Program, Stanford University, Stanford, CA

Abstract

Black Americans with chronic kidney disease (CKD) progress faster than Whites to end-stage renal disease. A variant in the apolipoprotein L, 1 (ApoL1) gene is associated with faster-progressing CKD, explaining most of the observed difference. The variant leads to CKD sub-types with distinct, shared phenotypes that are sometimes recognized by clinicians, who tend to assign diagnoses of hypertensive nephropathy. Using these insights, the objective was to identify patients with ApoL1-like CKD using only clinical data.

Introduction

Twenty-six million Americans are currently diagnosed with chronic kidney disease (CKD) and, amongst those affected, Blacks are approximately twice as likely as Whites to progress to end-stage renal disease (ESRD)1. The results of recent GWAS show that Blacks homozygous for a missense mutation in the apolipoprotein L, 1 (ApoL1) gene have a 10- to 29-fold greater risk for CKD1. This ApoL1-associated CKD is of a spectrum of three hypertension-related sub-types, with shared, distinct phenotypes2, and explains most of the increased rate of ESRD in Blacks. The variant is found almost exclusively in those of recent West African ancestry. The co-occurrence of hypertension with this sub-type of CKD often leads to an unspecific clinical diagnosis of hypertensive nephropathy.

Given the above findings, the aims of the present study were to: (1.) combine known phenotype similarity, diagnosis patterns and disease prevalence of ApoL1-associated CKD to identify a cohort of Black patients with putative cases, using only clinical data; (2.) compare results with previous findings obtained from patients with known ApoL1 risk-variant (Chronic Renal Insufficiency Cohort).

Methods

The method presented here (1.) uses hierarchical clustering of a standardized matrix of lab values to identify clinically similar CKD patients, (2.) picks the highest-rank cluster by frequency of patients with a diagnosis of hypertensive nephropathy (putative ApoL1 high-risk cohort), and (3.) constructs a control cohort (putative ApoL1 low-risk cohort) matched on time since CKD diagnosis of patients appearing in the top-ranked cluster from step (2.). The method was run on two separate datasets (primary and validation) and a combined dataset.

Results

Similar differences in rate of annual decline in kidney function (as measured by estimated glomerular filtration rate) were observed in putative ApoL1 low- and high-risk Black patients in primary, validation, and combined datasets (combined dataset: -0.4±2.8 ml/min/1.73m² low-risk vs. -3.0±5.8 high-risk; t-test, p=0.03) but were equivalent when applying the method to a control cohort of White patients (-1.2±2.8 vs. -1.6±2.5; t-test, p=0.15). By comparing the clinical profiles of both Black cohorts, the observed results were not explained by baseline differences in kidney function, time since diagnosis of CKD, blood pressure, age, gender, or diagnostic history of hypertension. Similar rates of decline in kidney function were observed in the Chronic Renal Insufficiency Cohort.

Conclusion

Even without known genotype, the method repeatedly identified a cohort of patients with rate of annual decline in kidney function consistent with a previous study using patients with known ApoL1 genotype. Furthermore, the results help disambiguate the role of hypertension as a cause of faster-progressing CKD observed in Black patients.

References

A Platform for Generating and Validating Breast Risk Models from Clinical Data: Towards Patient-Centered Risk Stratified Screening

Nova F. Smedley1,3, Ngan Chau2, Antonia Petruse2, Alex A.T. Bui PhD3, Arash Naeim MD PhD2, William Hsu PhD3
1Department of Bioengineering; 2Athena Breast Health Network, Jonsson Comprehensive Cancer Center; 3Medical Imaging Informatics Group, Department of Radiological Sciences, University of California, Los Angeles, CA

Introduction. Breast cancer prevention is often the center of great debate as to who should be screened and whether current screening policies are effective. In 2005, the Institute of Medicine advocated integration of technology, biology, and risk stratification for the development of screening models in breast cancer, noting that personalization could improve the predictive value of the screening test and ensuing interventions. Last year, the Nelson et al. enumerated 13 general risk models and 50 studies on BRCA prevalence to voice the need to update breast cancer screening recommendations. Today’s current risk models do not adequately reflect the increasing importance of quantitative imaging and gene expression as biomarkers for disease prediction. This work presents an informatics platform that integrates multimodal data sources (demographics, clinical, imaging, genomic, outcomes) with statistical tools to validate and generate risk models in a cohesive manner using observational clinical data collected on screening patients. The platform addresses challenges related to managing and processing this growing repository of patient data for the purposes of personalized risk-stratification.

Methods. With institutional review board approval, we developed a relational database tied to a high-performance cluster (HPC) and cloud storage solution that was Health Insurance Portability and Accountability Act (HIPAA) compliant. Our initial work focused on 1,401 patients enrolled through the Athena Breast Health Network, which was recently funded to study the effectiveness of personalized screening regimens. Patients completed a questionnaire that asked about breast cancer related risks (e.g., family history, hormonal factors). Imaging-derived information (e.g., BI-RADS scores), diagnostic tests (e.g., biopsies, molecular analysis), and other clinical information (e.g., comorbidities) were extracted from electronic health records (EHR). Initially, patients were mapped to the most common risk model for breast cancer screening and diagnosis: the 1989 Gail model with subsequent refinements. The model’s risk factors have been incorporated or found in other subsequent popularized models, demonstrating their validity over time in measuring the relative risk of an individual developing breast cancer. The HPC provided a centralized location for processing raw patient data and run preprocessing steps such as sequence alignment (for patients with biospecimens collected) and breast density quantification (from unprocessed mammography images). Whenever possible, data from the EHR were used to infer missing information needed to instantiate these models. Risks were validated by stratifying patients in terms of their clinical indication of where they fall along the diagnosis guidelines. To generate improved models, statistical analysis (e.g., logistic regression) can be performed using the captured data to further formulate models that incorporate novel features (e.g., gene expression, quantitative breast density); these models can then be validated using data on the population to determine whether additional features improve predictive performance.

Discussion. We present preliminary work on establishing an informatics platform that leverages high performance computing infrastructure at our institution to manage and process protected health information. Our initial efforts have focused on breast screening and improving risk stratification models, but this platform will be generalized to other clinical domains (e.g., other cancers, neurological diseases) as well. Using the platform, we identified that 36% of patients with a Gail score of over 2 had a confirmed diagnosis of cancer; our initial experiences in establishing this infrastructure has enabled us to quickly create and validate improved models of individualized risk.

References
Systemic Risk Analysis for Use Cases for Safety-Related Usability of EHRs

Michael W. Smith, PhD¹, Daniel R. Murphy, MD, MBA¹, Dean F. Sittig, PhD², Elise Russo, MPH¹, Hardeep Singh, MD, MPH¹
¹Michael E. DeBakey VA Medical Center and Baylor College of Medicine, Houston, TX
²University of Texas Health Science Center, Houston, TX

Introduction

Usability is critical to electronic health record (EHR) safety. However, there is little guidance on what use cases to target in a usability evaluation for improving EHR-related safety. While basic use cases have been proposed, they do not incorporate how user interactions with an EHR are affected by multiple socio-technical dimensions of the EHR-enabled healthcare delivery system. Targeting all the different permutations of user persona, clinical content, hardware, and workflow would be impossible. However, newer risk analysis methods, developed for application in complex socio-technical systems, may offer a means for informing the development of use cases for evaluating usability for EHR-related safety. The Functional Resonance Analysis Method (FRAM)¹ focuses on variability in the performance of a function (i.e., an action, such as ordering a test, that produces an outcome necessary to achieve a goal) and how it can interact with variability from other coupled functions. This can lead to conditions that exceed what can be managed safely. The aim of this analysis is to use FRAM to study the functions (and their interactions) involved in follow-up of abnormal test results in outpatient care settings using EHRs, in order to inform the development of key use cases to detect usability issues impacting safety.

Methods

We interviewed a total of 50 staff involved with testing processes at three large clinics, including lab, radiology, IT, and primary care personnel. From these interviews, we identified functions and factors directly related (“core functions”) or indirectly related (“support functions”) to abnormal test follow-up. We then developed the FRAM model in a stepwise process. We assessing the potential variability encountered in each function using an 8-dimension socio-technical model to encompass different sources of variance in performance. For example, the performance of the function Order test can vary in duration and risk of error depending on the complexity of the patient and the EHR training of the provider. To identify elements for inclusion in use cases for evaluation for EHR safety, we looked for functions and sources of variance that had impact throughout the full set of functions.

Results

Upstream functions influenced the timeliness and reliability of test result follow-up. In primary care, the number and complexity of patients influenced the time used for three core functions: Order test, Obtain test sample/radiology images, and Find test results in EHR inbox. They also constrained the support function Maintain time for administrative tasks, which served as a resource for two core functions, Find test results in EHR inbox and Order follow-up, and for two support functions, Detect and fix problems in test orders, and Maintain training on EHR (which itself is a resource for Order test and Find test results in EHR inbox). In the area of diagnostic services, patient number and complexity also affected Ensure the functioning of lab & imaging capabilities, which is a resource for Update list of test order options in EHR and Detect and fix problems in test orders. These findings indicate that time pressure and number of tests are key factors affecting performance throughout the process. The functions Detect and fix problems in test orders, Maintain training on EHR, and Update list of test order options in EHR compete with time demands but can save time in the long run. Based on these results, example use cases with a safety focus that encompass these elements could include: 1) The user must detect a problem with a test order (e.g., out of date orderable item) under conditions of time pressure and a large volume of orders; and 2) The user must order a test under conditions of minimal training and a list of order options that includes some out-of-date choices.

Conclusion

FRAM can be used to identify functions and factors that are important for safety of abnormal test result follow-up in EHR-enabled healthcare systems, and can inform key use cases to detect important safety-related usability issues.

References

Prediction of Colorectal Surgical Site Infections Using Risk Factors

Sunghwan Sohn, PhD1,2, Majid Rastegar-Mojarad, MS1,2,6, James M. Naessens, ScD1,3, Elizabeth B. Habermann, PhD1,3, David W. Larson, MD4,5, Hongfang Liu, PhD1,2
1Department of Health Sciences Research, 2Biomedical Statistics and Informatics, 3Health Care Policy and Research, 4Department of Surgery, 5Division of Colorectal Surgery, Mayo Clinic, Rochester, MN; 6Department of Health Informatics and Administration, UW-Milwaukee, Milwaukee, WI

Introduction

Surgical site infection (SSI) is the most common healthcare-associated infection, accounting for 31% of all hospital inpatient infections and resulting in increased length of hospital stay and cost. The patient characteristics in National Surgical Quality Improvement Program (NSQIP) can be regarded as potential factors in predicting SSI. In this study, we analyzed perioperative risk factors from both NSQIP data and clinical notes and further employed a Bayesian network classifier to investigate the prediction capability of SSIs using those risk factors.

Materials and Methods

We used NSQIP data from colorectal surgery (CRS) performed at the Mayo Clinic Methodist hospital from 2010 to 2012. This consisted of 751 surgery cases with 67 SSIs (i.e., 36 superficial incisional SSIs, 5 deep incisional SSIs, and 26 organ/space SSIs). A Bayesian network was developed to detect SSIs using risk factors in both NSQIP data and clinical notes. The SSI risk factors we used are: 1) pre-operation: age, gender, smoking status, pre-op hospital stay, diabetes, anemia†, BMI, steroid use, blood transfusion, prior operation; 2) intra-operation: ASA score, operation type, wound classification, duration of operation, laparoscopic vs others; 3) post-operation: anemia†, ascites, wound disruption, abdominal tenderness†, antibiotics†, fever† († denotes variables extracted from clinical notes using natural language processing; the other variables are from NSQIP data).

Results

Table 1 shows the ROC area under the curve (AUC) of the Bayesian network classifier using different sets of variables with 10-fold cross validation. For comparison, a logistic regression was also employed (corresponding ROC curves are shown in Figure 1). Both methods produced relatively high specificity but low sensitivity. The basic statistics and odd ratios of patient characteristics were also examined, and these are described in Table 2.

Table 1. Bayesian network AUC for SSI prediction

<table>
<thead>
<tr>
<th>Feature used</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre/intra-op risk factors</td>
<td>0.643</td>
</tr>
<tr>
<td>risk factors</td>
<td>0.721</td>
</tr>
<tr>
<td>major five variables†</td>
<td>0.609</td>
</tr>
</tbody>
</table>

† variables in Table 2

Table 2. Major risk factors for colorectal SSI

<table>
<thead>
<tr>
<th>Patient factor</th>
<th>SSI (N=67)</th>
<th>No SSI (N=684)</th>
<th>Odd Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI (&gt;=30)</td>
<td>37.3% (25)</td>
<td>26% (178)</td>
<td>1.69</td>
</tr>
<tr>
<td>Diabetes</td>
<td>19.4% (13)</td>
<td>8.6% (59)</td>
<td>2.55</td>
</tr>
<tr>
<td>Postop anemia</td>
<td>40.3% (27)</td>
<td>20.3% (139)</td>
<td>2.65</td>
</tr>
<tr>
<td>Preop steroid</td>
<td>16.4% (11)</td>
<td>12.7% (87)</td>
<td>2.55</td>
</tr>
<tr>
<td>Blood transfusion</td>
<td>4.5% (3)</td>
<td>2.8% (19)</td>
<td>1.64</td>
</tr>
</tbody>
</table>

( ) denotes number of cases; % is computed by number in ( ) / N

Discussion

Certain patient characteristics can be used as risk factors to predict SSIs. A Bayesian network performed better than a traditional logistic regression for the prediction. For real-time SSIs surveillance those risk factors can be utilized supportively along with SSI indicative keywords from clinical narratives to accurately monitor SSIs.
Representing and Validating Cancer Study Metadata Standard Using RDF Shapes Expression Language

Harold R. Solbrig¹, Eric Prud’hommeaux², Christopher G. Chute³, Guoqian Jiang¹
¹ Department of Health Sciences Research, Mayo Clinic, Rochester, MN; ² W3C/MIT, Boston, MA; ³ Johns Hopkins University, Baltimore, MD

Abstract

The RDF Data Shapes Working Group is developing a W3C Recommendation for the Shapes Constraint Language (SHACL), which formally describes structural constraints on instances in Resource Description Framework (RDF) graphs. The objective of this study is to create a publicly available SHACL-based schema that defines a common way to represent the subset of the ISO/IEC 11179 Metadata Repository data element model in RDF. The goal of this project is to be able to produce a shared and extensible set of tools for validating and harmonizing data element definitions within the RDF / OWL environment and, in the longer term to use these harmonized definitions to validate, transform and integrate disparate cancer study data into cohesive and consistent repository.

Introduction

ISO/IEC 11179 Common Data Elements (CDEs) are used to support the semantic interoperability and aggregation of research data from disparate sources¹. The National Cancer Institute (NCI) has developed a cancer study data standard repository (caDSR)¹ based on the ISO/IEC 11179 metadata registry standard² that is used to describe and annotate CDEs in the cancer study domain. We have been developing a Resource Description Framework (RDF) triple store in combination with a Web Ontology Language (OWL) reasoner to validate CDE definitions and integrate them across domains. While this is producing promising results on a local level, the real advantages of the approach will come from scalability -- the ability to apply and extend tools for use across multiple domains and situations.

The Resource Description Framework (RDF) Shape Constraint Language (SHACL)³ provides a number of interesting opportunities along this front, including:

- Publishing and sharing the RDF model used in a given ISO 11179 compliant RDF Data Set or Triple Store;
- The ability to validate and transform external data into a shared structure and format;
- The ability to transform the definitions of the ISO 11179 Data Element definitions into SHACL Shape Constraints for use in validating and integrating the target data.

Methods and Results

In this pilot study, we focused on core data meta-model defined in Part 3 of the ISO/IEC 11179 standard², including definitions of: Data Element, Data Element Concept, Object Class, Property, Value Domain, Permissible Value, Conceptual Domain and Value Meaning. We used a simple representation of SHACL called as Shape Expression (ShEx)⁴ to define the target meta-model attributes. We then used these definitions in combination with a ShEx based validation tool to drive and verify the transformation of a subset of the NCI caDSR Common Data Elements from XML to RDF. Our next steps will include: 1) A tool that will allow us to generate ShEx rules that describe the instance data described in the NCI caDSR repository. 2) A tool to transform UML models into SHACL equivalents. In summary, the pilot study provides useful insight into the strength of the SHACL-based framework for representing and validating cancer study metadata standards.

Acknowledgments: This work has been supported in part by funding from caCDE-QA (1U01CA180940-01A1).

References

PubMed ‘Early Alerts’: A Pilot Study to Support Prospective Detection of Emerging Adverse Drug Events

Alfred Sorbello1, DO, MPH; Anna Ripple2, MLS; Olivier Bodenreider2, MD, PhD  
1US Food and Drug Administration, Center for Drug Evaluation and Research, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA. 2Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA.

Background: The FDA traditionally monitors the safety of marketed drugs by analyzing reports submitted by manufacturers and consumers to the FDA Adverse Event Reporting System (FAERS). In order to enhance prospective detection of emerging adverse drug events (ADE), we investigated leveraging existing PubMed “MyNCBI” functionalities and searching resources to survey the biomedical literature for the latest published safety information in the use case of the new oral hepatitis C drugs.

Methods: Using PubMed “MyNCBI” cubby functionality, we established a search strategy designed to retrieve citations most recently added to PubMed and provide automated weekly emails (PubMed ‘Early Alerts’) with abstracts and links to available full text citations. The search strategy (Table 1) includes three criteria: (1) Drugs of interest; (2) Citation scope: citations from MEDLINE journals, as soon as they are entered in PubMed (prior to indexing); (3) Time Period: week of interest.

<table>
<thead>
<tr>
<th>(“drug A”[all fields] OR &quot;drug B”[all fields] OR …)</th>
<th>Drugs of interest conjoined with the Boolean operator OR and searched in all fields including the Title, Abstract, and other terms (including Author Keywords) fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND</td>
<td></td>
</tr>
<tr>
<td>((publisher[sb] NOT pubstatusnihms[All Fields] NOT pubstatuspmcsd[All Fields] NOT pmcbook[All Fields]) OR inprocess[sb])</td>
<td>Includes citations supplied by publishers that are not yet indexed and have not completed quality control, and includes citations in the review process; Excludes citations from non-MEDLINE journals (author manuscript or publisher-supplied citation submitted to PMC under the NIH Public Access Policy, book or book chapter citations available on NCBI bookshelf)</td>
</tr>
<tr>
<td>AND</td>
<td></td>
</tr>
<tr>
<td>“2015/Month/Day 15.00” [MHDA]:&quot;2015/Month/ 7 Days Later 15.00” [MHDA])</td>
<td>Restricts the search to a 7-day time period; MHDA is automatically added by PubMed for each weekly search—this is one of the parameters that can be set when stored searches are set up in the NCBI cubby.</td>
</tr>
</tbody>
</table>

Results: A typical weekly PubMed ‘Early Alerts’ email delivery consists of 12-15 citations from domestic and international journals spanning 5-8 oral hepatitis C drugs of interest. The citations reflect human, animal, and in vitro research from the US and other countries with varied publication types, including reviews, letters, clinical trial reports, pharmacokinetic studies, case reports, and observational studies. Our search strategy is designed for recall. Its precision can be increased significantly by requiring the presence of specific words in the titles or abstracts (“safety”, “toxicity”, “adverse”, and “tolerability”).

Conclusions: By leveraging existing PubMed “MyNCBI” functionalities and searching resources, we can more efficiently canvass across a broad range of journals and publication types in PubMed for the latest drug safety data. This approach lessens reliance on time-consuming and often inefficient ad hoc searches, and it complements traditional approaches to finding relevant safety information on drugs of interest. Evaluation by FDA Medical Officers and Safety Evaluators indicates that the PubMed ‘Early Alerts’ complements their efforts to discover emerging ADEs as part of their regulatory review activities.

Acknowledgements: Funding support received from the FDA/CDER Critical Path Program and the Intramural Research Program, NIH, National Library of Medicine. Disclaimer: The views expressed are those of the authors and do not necessarily represent the views of the US FDA, the NIH, or the US Government.
Annotating ADLs and IADLs in Veterans Affairs Clinical Documents

Brett R South, MS, PhD1,2, Danielle L Mowery, PhD1,2, Lee Christensen, MS1,2, Adi V Gundlapalli, MD, PhD1,2, Melissa Tharp, BS2, Marzieh Vali, MS3, Marjorie Carter, MS2, Mike Conway, PhD2, Salomeh Keyhani, MD3, Wendy W Chapman, PhD1,2
1VA Salt Lake City VA Health Care System, and 2University of Utah, Department of Biomedical Informatics, Salt Lake City, UT, 3San Francisco VA Medical Center, San Francisco, CA

Abstract Descriptions of six activities of daily living (ADLs) and seven instrumental activities of daily living (IADLs) are documented in unstructured U.S. Department of Veterans Affairs (VA) healthcare data. Natural Language Processing (NLP) can be used to extract this information from clinical free-texts to predict the likelihood of nursing home admissions after inpatient admission or as measures of functional status in risk adjustment. We assess annotation guidelines for ADLs/IADLs used to build a reference standard for NLP algorithm development and report descriptive statistics describing prevalence, inter-annotator agreement, content structures, and mention prevalence for explicitly described functional impairment.

Methods We constructed a cohort of 1,000 patients admitted for congestive heart failure, acute myocardial infarction, pneumonia, and stroke and extracted their clinical documents (n=353,889) for calendar year 2012 extracted from the VA’s Corporate Data Warehouse. We randomly sampled 180 documents from a physician-reviewed document set (n=52,304) targeting those documents most likely to contain mentions of ADLs/IADLs: Nursing Assessments, Social Work consultations, Occupational/Physical Therapy, ER Physician, Primary Care, H&P, Critical Care, and Discharge Summaries. Using eHOST1, two reviewers independently annotated mentions of ADLs/IADLs: a third reviewer arbitrated disagreements creating a reference standard using a previously developed schema and guideline2. We conducted post hoc analysis to classify mentions into ADL subcategories (Bathing/Hygiene, Dressing, Eating, Toileting, Transferring, Walking) and IADLs (Communication, Managing finances, Managing Healthcare, Home Maintenance, Preparing meals, Shopping, Transportation). We determined whether the annotated mentions describe functional impairment and where implementation of an algorithm like ConText3 would be warranted for classifying impairment. We report inter-annotator agreement, ADL/IADL prevalence, content structures (i.e. semi-structured input vs narrative), and mention prevalence for descriptions of functional impairment.

Results In the text set of 180 documents, a total of 1,490 (8.7 per document) for ADLs/IADLs were annotated. Inter-annotator agreement was 71% (substantial agreement) for annotation of ADL/IADLs. The bulk of these observed annotations (800, 54%) were ADLs classified into Bathing 74 (9%), Dressing 36 (4.4%), Eating 75 (9.1%), Toileting 175 (23.7%), Transferring 81 (9.9%), and Walking 245 (29.8%). IADLs represented only a small number of annotations containing only five of the seven categories including Managing Finances 2 (<1%), Managing Healthcare 2 (<1%), Home Maintenance 12 (1.5%), Preparing meals 3 (<1%), and Shopping 2 (<1%). Only 22% were generated from narrative descriptions; the majority were components of semi-structured text input, such as tables or question/answering templates, which are not easily classified with standard NLP techniques. The majority of ADLs/IADLs did not imply functional impairment - only 177 (21%) explicitly stated the patient was functionally impaired for that given category.

Conclusions Additional work will involve developing rules to deal with situations where the ADL/IADL is mentioned as part of semi-structured text or where additional contextual information is required to determine functional impairment. Developing NLP algorithms that reliably extract ADLs/IADLs scaled to large patient cohorts will pose challenges but will produce useful predictors of healthcare outcomes.

References
Inexpensive Radio Communications System for Wheelchair Users

James R. Sprague, Dorothy Curtis, MIT/CSAIL, Cambridge, MA, USA

Abstract
Wheelchair-bound residents of assisted living facilities sometimes require assistance from staff when they are isolated or away from existing call buttons. Those with limited manual dexterity could use this system when their wheelchair slips off the sidewalk. Getting help in this situation is especially important to users with advanced multiple sclerosis, who are at risk when exposed to excessive heat. This project conducted a Proof-of-Concept trial using two-way radios to allow residents of an assisted living facility to communicate with staff members from their wheelchairs. The radios were effective and reliable on the facility campus. For certain subsets of this population they appear to be a helpful and cost-effective assistive technology.

Overview

The Boston Home (TBH), based in Dorchester, Massachusetts, is a not-for-profit specialized care residence for adults with advanced Multiple Sclerosis and other neurological diseases. 4 wheelchair-bound residents volunteered to have their chairs equipped with a Family Radio Service (FRS) band radio in a standardized enclosure with a custom-mounted microphone, speaker, and Push-to-Talk switch. A staff member carried a standard FRS radio during normal business hours. All communications with this system were broadcast over an open channel which could be heard by anyone in the trial or anyone else nearby with proper equipment, so there was no expectation of privacy. Residents were notified of privacy issues and then allowed and encouraged to use the system as they saw fit. All uses of the system were centrally recorded and later reviewed and tallied. Exit interviews were conducted at the end of the 6 week trial.

Trial Results

Post-trial analysis of the trial confirmed that there were no uses of the system to address urgent needs, though no participants reported an urgent situation where the system failed. All users reported having temporarily muted the radios during activities classes or other group activities, but none of them requested that the radios be removed before the end of the trial. All participants felt that the system could be useful for urgent communications, especially in summer months when they are more likely to be outside on the facility grounds.

Analysis of the recordings also showed that most uses of the system were for confirming/testing the operation of the system, combined with social interactions with the staff. The system was used on occasion to locate individuals or to schedule visits to the wheelchair repair facility. Over the course of the trial, daily uses dropped off substantially from a one-time peak of 20 to less than one per day on average. This is attributed to the waning novelty of the system over time.

Conclusions

The trial was considered a success as a Proof-of-Concept. Each radio required less than $100 in parts and required no ongoing licensing fees or external infrastructure (e.g. WiFi). This can be compared to $20 for a low cost cell phone + $40/year for service, assuming the user has the dexterity to use the cell phone. (Interested readers are referred to Technology Assessment of Communication Systems for Wheelchair Users in an Assisted Living Setting for more information on comparative costs). The radios worked reliably across the TBH campus (approximately 500’x500’). Participants appeared to enjoy having direct access to the staff, and believed that the radios could be useful in an urgent situation. In its current form, this system would be unsuitable to those with extensive speech, hearing, or upper body motor deficits, though some enhancements have been proposed that would expand the target population.
Machine-to-Machine (M2M) Communication in Home-care
Basile Spyropoulos, PhD, Aris Tzavaras, PhD and Maria Botsivaly, MSc.
Technological Educational Institute (TEI) of Athens, Athens, Greece

Abstract
The purpose of this work-in-progress paper is to present the status of our attempts to utilize "hybrid" Machine-to-Machine (M2M) Communication, in Home-Healthcare. M2M communication and the Internet of Things (IOT) are growing quickly and we consider that they will be soon adopted in Home-Healthcare and Point-of-Care testing systems, because of their advantages, concerning interoperability and automation of the said systems.

Introduction
There are a lot of emerging technologies during the last five years that allow for “Machine Type Communication, MTC” leading to a new “Internet of Things, IOT”, enabling “Machine-Cooperation” without human intervention.

Methods and Resources
The purpose of this work-in-progress paper is to present the status of our attempts to utilize “hybrid” M2M-Communication, in Home-Healthcare. We use the term “hybrid systems” since the systems under development, must comprise of, due to their Healthcare related tasks, both, human mediated and machine-only “operation modes”. We are presently focusing on the following Home-care tasks, as displayed in Figure 1 (right).

Results
We have adopted the approach and the terminology of the International Telecommunication Union (ITU) and more specific, the M2M service layer and architectural framework requirements [1], the Application Programming Interface (API) [2] and Protocols and the remote patient monitoring/assisted living requirements (RPM/AL) [3]. Data are transferred from sensors linked to the patient, forming a body area network (BAN), to a gateway that manages the sensors and it can be continuous or periodic, depending on the sensor and the Biosignal types. The acquired data are stored in a gateway and/or uploaded to a Medical Information System, located in a Wide Area Network. In the gateway, the data are linked to the patient demographics. Caregivers, such as Medical Doctors, Nurses etc. can access the Patient’s personal information, in accordance with privacy and security requirements.

Conclusion
We have described briefly our efforts to make use of ITU-compliant M2M-Communication in Home-care and in PoCT, taking into account the “state-of-art” presented in Figure 1 (left) and looking for the ICT-Standards’ Essential-Patents possibly involved [4]. The system is presently still under development and it is being tested by comparing the automatic performance of each function to the manual one.

References
Bringing Context to Data Analytics: A Hybrid Approach to Understanding Clinical Workflow

Bryan Steitz¹, Kim M. Unertl, PhD, MS²
¹The Pennsylvania State University, University Park, PA; ²Vanderbilt University School of Medicine, Nashville, TN

Abstract

Understanding workflow complexities is integral to achieving maximum effectiveness within clinical environments. Additionally, the coordination of patient care across multiple care providers has surfaced as an efficiency issue. Understanding workflow can help accommodate differences in user needs, which can improve HIT use and adoption. We offer an iterative, mixed methods, approach of workflow modeling to further the analysis and understanding of clinical workflow.

Introduction

Clinical care environments are complex systems supporting highly collaborative and dynamic conditions. Workflow evaluation has emerged as a method to analyze the complexity of clinical environments with relation to patient care coordination and identification of operational concerns. Quantitative workflow sources can provide some types of workflow-related data, but can also lead to correlation of events without understanding causation. Qualitative data provides insight into the interworkings of the clinical setting but presents challenges of scale, scope, and generalizability. We combined qualitative methods and data science approaches in iterative cycles to determine how researchers could build richer understanding of clinical workflow through a mixed methods approach.

Methods

We collected and analyzed both quantitative and qualitative data from three areas at the Vanderbilt Health One Hundred Oaks facility: the Vanderbilt Breast Center, an associated infusion clinic, and pharmacy. Qualitative data were collected through observation over more than 27 hours, divided across the three areas. During observations, free text notes about workflow were recorded. Resulting notes were coded and analyzed for common themes. Quantitative data were collected from Vanderbilt’s electronic outpatient whiteboard. The data set was assessed to identify and address incomplete cases and other data quality issues. Quantitative data were processed through timeline visualization tools created using JavaScript and D3 to determine themes. Iterative analysis of combined qualitative and quantitative themes was conducted after each day of observations.

Results and Discussion

Quantitative data collected between April 2014 and May 2014 provided patient flow data for 5592 unique patients across the two clinics. Initial findings indicate the breast center has the highest level of patient flow with over 2000 unique patients visiting the clinic over the two-month period. Though patient throughput was high, the breast center had a low rate of miscommunication with only seven coded instances. The infusion clinic indicated similar results, noting only five coded instances of miscommunication in 816 patient visits. This low rate of miscommunication – defined as an instance in which additional clarification or a break in workflow due to a lack of communication – may be related to the clinic’s high utilization of the electronic whiteboard. Patients spent the longest amount of time in infusion rooms, occupying the room for an average of 2 hours and 35 minutes per visit. This length of time spent in the infusion room is likely due to the time required to receive an infusion. However, upon arriving in the infusion room, an average of 29 minutes and 30 seconds elapsed before the infusion began. This additional wait time, while outside of the traditional waiting room, offers significant opportunity for efficiency improvements.

Conclusion

Our iterative mixed methods approach allowed us to provide a context to our quantitative data by understanding the inner workings of the clinical areas. Though data from the electronic outpatient whiteboard contributed to our unique understanding of the patient flow, data from employee time clock software, registration software, and pharmacy software could help to provide additional workflow metrics. We eventually envision our visualization tools being implemented into Vanderbilt Health’s electronic health record system to provide healthcare team members and clinic managers with a better understanding of the workflow in different clinics.
Using the Consolidated Framework for Implementation Research (CFIR) to Evaluate EHR Population Health Management Tools
Christina L. Stephan, MD PhD, Lori L. Popejoy PhD APRN GCNS-BC FAAN, David R. Mehr, MD, Esmaeel Rahmani, MD MSHI, Katie Wilkinson, MA, Eduardo J. Simoes, MD MSc DLSHTM MPH
1University of Missouri, Columbia, Missouri; 2Tiger Institute, Columbia, Missouri

Introduction
How electronic health record (EHR) innovation is implemented in clinical practice is influenced by the contextual environment. The Consolidated Framework for Implementation Research (CFIR) provides a comprehensive framework for analyzing an implementation process. We examined one component of a multi-faceted intervention to improve chronic disease care: an EHR-based provider panel dashboard, which shows performance on quality measures for a provider’s panel of patients. The objective of the study is to describe the use of the CFIR to analyze provider focus-group feedback regarding the implementation of the dashboard.

Methods
We conducted two focus groups to obtain information about barriers and facilitators that influenced physician adoption of the dashboard after the implementation. Focus groups of 10 and 6 physicians were conducted, recorded, and transcribed. The 39 elements in the CFIR were used to structure analysis of contextual barriers and facilitators to adoption of the dashboard. Results were used to describe opportunities for improved implementation.

Results
All of the focus group comments fit in the CFIR structure. Participants reported that the dashboard implementation was suboptimal. Three opportunities for improvement were identified: 1) Increasing provider engagement and communication early in the development and implementation process; 2) Increasing leadership engagement at the practice level to address work flow changes and clinical integration of new tools and practices; 3) Formalizing mechanisms for efficiently obtaining iterative feedback from clinical and administrative stakeholders to improve effectiveness of implementation efforts.

Discussion
Use of the CFIR has not yet been well documented in the informatics literature. The CFIR complements the established informatics socio-technical evaluation framework. To efficiently implement new tools or innovations, it is essential to understand the contextual factors that affect the implementation process and influence adoption. Use of the framework can also help stakeholders and investigators consider potential barriers that might be addressed prior to implementation. By systematizing the identification of barriers and facilitators, the comprehensive framework for implementation research enables a planning process which can accelerate adoption of innovations.

Conclusions
The CFIR provides a useful framework for understanding intervention implementation and opportunities for improving implementation processes. In this example of implementation of an EHR-based provider panel dashboard, issues of engagement, leadership, and feedback were identified as particularly important.

References
Matrix Completion Methods and Imputation for EMR-Based Prediction.

Erika Strandberg, MS\textsuperscript{1}, Mohsen Bayati, PhD\textsuperscript{2}

1. Stanford University Biomedical Informatics Training Program. 2. Stanford University Graduate School of Business.

Background
The availability of electronic medical records (EMRs) has allowed for the use of a large amount of clinical data for predicting healthcare outcomes; however, the high proportion of missing entries can make predictions and risk assessment difficult. Few studies have compared methodology for missing data completion with the objective of optimal predictive performance; even fewer studies have looked at imputing missing values in medical data with high amounts of incompleteness. In particular, lab results could be informative in determining patient health status; however, due to cost limitations, availability, risk and appropriateness, most labs are not performed on all patients. As such, only frequently performed labs are typically used in predictive modeling. In these situations, imputing lab results could be useful because (1) there might be a rarely performed test or combination of tests that is informative, (2) signal in repeated lab tests could be useful but not all patients have labs measured at the same intervals or (3) lab tests requested could be indicative of differential diagnosis so the act of performing or not performing a lab test could rule out other conditions. Since laboratory tests are typically performed when a practitioner suspects a certain diagnosis, missing values are not missing at random.

We compare the performance of existing data completion and imputation methods parameterized for predictive performance to novel imputation methods using both simulated data and real EMR data with hospital-associated conditions and mortality.

Methods
Simulated data was obtained by sampling random (patient, lab) pairs where the marginal distribution of labs matched the empirical distribution of labs from a large urban hospital’s EMR. Response variables determined by logistic regression on a random subset of lab results. Lab results were taken from the first day of admission to the ICU from the MIMIC II ICU database\textsuperscript{1}. The outcomes of interest were hospital-associated venous thromboembolism and mortality.

We imputed missing values using mean and k-nearest neighbors (kNN) imputation. We also used SoftImpute, a matrix completion method successful on the Netflix data set\textsuperscript{2}. Finally, we implemented novel aggregated multiple imputation method (MI) and a feature selection SoftImpute (FSSI), which performs feature selection prior to imputing using SoftImpute. Parameters were chosen through 10-fold cross validation to maximize area under the receiver operator characteristic curve (AUC) on a training set. Logistic regression with a lasso penalty was used to perform feature selection and classification and overall performance was assessed on held-out test sets.

Results
SoftImpute, FSSI and kNN all performed well for the different feature sets and classification tasks assessed. MI did not exhibit consistent performance. There was no clear best performer over all classification tasks; however, when the number of patients is high, kNN is not computationally tractable.

Conclusion
Results from this study indicated that the choice to impute is a good one; however, the best method to use may not be the same for all tasks. FSSI holds much promise due to its speed and performance.

References
Clinical Efficacy of Virtual Reality Rehabilitation Intervention for Patients with Low-back Pain: A Randomized Controlled Trial

Wu-Chen Su, MS1, Shih-Ching Yeh, PhD2,3, Si-Huei Lee, MD4, Hsiang-Chun Huang, MS3
1Chang Gung University, Taoyuan, Taiwan; 2Sun Yat-sen University, Guangzhou, P.R. China; 3National Central University, Taoyuan, Taiwan; 4Taipei Veterans General Hospital, Taipei, Taiwan

Introduction
Given that eighty percent of all people may have experienced low back pain (LBP) during their lifetime1 this health condition is a very common globally. Furthermore, LBP rehabilitation requires a considerably long commitment of time which in turn may cause patients to lack motivation to finish their entire course of treatment. As a result, LBP poses a substantial impact on society (e.g., increased medication cost). Recently, entertaining virtual reality-based systems (e.g., Wii Fit Yoga) have been used to encourage patients to complete their LBP rehabilitation tasks in an effective way2. However, these commercial exercise programs do not offer the capacity for patients to incorporate learning patterns into daily activities to ease cumulative low back loads in their daily lives and to prevent the recurrence of LBP. Therefore, these limitations merited the development of a previously proposed VR-based approach3 that tracks and measures patients’ movements during rehabilitation treatment in real time, provides immediate visual and auditory feedback to users, and allows physiotherapists use statistical information tracked and analyzed for treatment supports (Figure 1). In this study, the authors further evaluate the undertaking of patients’ rehabilitation intervention through a randomized controlled trial to examine the difference in efficacy between the proposed approach and usual LBP care.

Methods
Forty patients were recruited and randomly assigned into either an experiment group (EG) or a control group (CG). The difference of duration of pain (mean ± SD months) between groups was not significant (EG: 12.3±9.5 months; CG: 15.7±8.7 months; p>0.05). Physical therapies were provided for both groups firstly (e.g., hot pack). Subsequently, the EG used the proposed VR-based system as a therapeutic medium, and the CG was treated using the usual LBP exercise programs with health education. The rehabilitation sessions were performed 2-3 times a week for 4-6 weeks, and each session was 60 minutes. Patients answered survey questionnaires before and after the interventions relating the scale of their pain intensity (0-10), the degree of unpleasantness (0-10)4 and the level of disability (0-100)4 in their daily life. The Mann-Whitney U-test was used for statistical analysis (significance level=0.05).

Results
The EG demonstrated significant improvement (p<0.05) via alleviation of pain intensity (5.1±2.6 to 1.1±0.9) and unpleasantness (4.1±1.6 to 1.4±1.2) compared to the CG (pain intensity: 5.3±2.3 to 4.3±1.2; unpleasantness: 4.3±1.2 to 3.3±2.2) while performing general daily life activities (e.g., walking, sitting, etc.). Additionally, the average pain intensity scores and mean unpleasantness scores at baseline were 8.1±1.6 and 9.1±0.5 for the EG, and 8.3±3.1 and 8.3±1.1 for the CG during the period of exercise. VR-based treatments yielded significant reductions in pain intensity and unpleasantness (p<0.05) at the end of the study, with scores of 2.1±1.6 and 1.4±1.2 for the EG, and 5.6±3.2 and 4.3±3.1 for the CG. In addition, the disability impact of LBP on the participants’ daily life of the EG decreased significantly (79.2±11.6 to 21.1±9.6) compared to that of the CG (78±13.1 to 53.5±11.6) after treatments (p<0.05).

Conclusions
The experimental results demonstrate that the patients of the EG who received the treatments utilizing the VR-based therapies have better physical and mental health improvements (e.g., sleeping, walking, etc.) than the patients of the CG receiving the usual care. Thus, the physiotherapists suggest this approach may effectively reduce the burden of LBP patients and can be used to re-align daily life behaviors to prevent recurrence of LBP. In addition, the functionality of this system can be extended to provide tele-rehabilitation services to the general public in the future.

References
Comparison of Patient Portal Usage between Employees and Non-Employees

Lina M. Sulieman1, Dara L. Eckerle Mize, MD1,2, Daniel Fabbri, PhD1,3, S. Trent Rosenbloom MD MPH1

Department of Biomedical Informatics, 2 Division of Diabetes, Endocrinology & Metabolism, Department of Computer Science; Vanderbilt University School of Medicine, Nashville, TN

Introduction

Vanderbilt University Medical Center (VUMC) policies encourage employees with job-related electronic health record (EHR) access to review their own record. We hypothesized that with easy access to their own EHR, employees would be less likely to use the tethered patient portal. We compared patient portal usage by employee and non-employee patients. Identified variations in patient portal usage and patterns will guide the design of new functionalities in the patient portal and policies around its use.

Methods

VUMC is a private, non-profit academic medical center in Nashville, Tennessee that has in place a locally developed EHR and Patient Portal. Core portal functions include secure messaging, access to some EHR data, appointment scheduling and bill management. Subjects included all VUMC employee and non-employee patients who accessed their portal account between August 2013 and September 2014. We obtained data from the Vanderbilt Research Database (RD), a database of clinical and related data derived from VUMC’s clinical systems. We collected information describing the frequency and types of actions users took while viewing their patient portal.

Results

During the study period, 17,917 employees accessed the EHR. A total of 12,044 (67%) of those employees logged into their own portal account. There were 95,808 portal users, 87% are non-employee patients. Individuals in both groups accessed their account a mean 10 and a median 5 times over the study period. The top ten most commonly performed actions were the same for employee and non-employee patients (Figure 1). Figure 2 shows that a greater percentage of employee versus non-employee patients used portal messaging. The two groups equally used functions to view EHR data such as lab results, radiology reports, visit summaries and problem lists.

Conclusion

Functionalities related to secure messaging are the most commonly used among both user groups. A greater percentage of employee versus non-employee patients use the secure messaging functionality of the patient portal to communicate with their provider. This may reflect greater awareness of the tool. Use of other patient portal functions did not differ between the groups. Analysis of how employees use job-related access to view their own EHR may guide the development of new tools for the patient portal.
Personal Health Information Management Strategies:
Experiences of Patients in the US and China

Si Sun, MS, Nicholas J. Belkin, PhD
Rutgers University, New Brunswick, NJ 08901

Background
Harnessing personal health information has become increasingly important for people living with chronic conditions. Research has found that health information can help patients improve their health care outcomes by enhancing motivations and help with problem solving and interpersonal relationship management. To effectively use health information, many patients developed information management strategies that are continuously improved through trial and error. Understanding these strategies can provide important guidance to the development of supporting technologies.

In this study, we explore the personal health information management (PHIM) strategies of people living with diabetes mellitus. We examine how these strategies differ between people from two distinct social contexts and living with different health conditions (i.e., type 1 or type 2 diabetes). Understanding the PHIM strategies with these differences in mind can provide insight into the translation of health care interventions and the adoption of health information technologies across contexts.

Methods
In this study, we conduct semi-structured one-on-one interviews with 60 diabetes patients (30 from the US and 30 from China). Both type 1 and type 2 diabetes patients are included. All participants are recruited through snowball sampling. The first author recruit the first few participants from her social network and ask the exiting participants to introduce more patients. Other recruiting methods such as distributing advertisements at community centers, hospitals, online health communities, and through email listservs did not successfully recruit participants. To address the issue of sampling bias posed by snowball sampling, we will conduct a survey with a larger sample size, as part of a larger study. Participants from the US are mainly recruited in the New York City area, while those from China are all affiliated with a large research university in southern China. Long distance interviews in the US are carried out over Skype or the telephone. Interviews in China are all conducted face-to-face. We also collaborate with participants to perform photo-documentation. After the first interview, the researcher and the participants collaborate to take photographs depicting the PHIM strategies.

So far, we have completed the interviews and photo-documentation with all 30 (one with type 1 and 29 with type 2) participants in China and 23 (twenty with type 1 and 3 with type 2) in the US. The poster will present a collection of PHIM strategies from 60 participants. We will report these different strategies in relation to participants’ health condition, social contexts, age, sex and occupation.

Preliminary Results
Patients maintain a wide variety of PHIM strategies, some are so ingrained in their daily routines that they become invisible to patients. Some examples of the strategies we identified so far include: (a) Extensive recording and integration strategy: Some patients keep an extensive record health information and integrate it across information items to extract theories about their condition and take actions accordingly. (b) Location-based integration strategy: Many patients keep all diabetes-related information and information management tools at a single location (e.g., a box) and always go to that location when performing self-management activities. (c) Location-based redundancy strategy: Patients may also utilize a distributed information system comprised of synchronized information technologies and duplicated documents and items for reminder and fail-safe purposes. (d) Leadership responsibility strategy: Several patients assumed leadership roles in patient support groups and relevant health care organizations, and have become information gatekeepers. (e) Collaborated improvement strategy: Some patients share their information with a network of family and friends who provide constant feedback to their performance and assist with regimen adjustments.

References

1708
Data-driven knowledge base evaluation: Translating an adult CDS tool for use in pediatric care

Katherine A. Sward, PhD RN¹, Christopher J Newth, MD FRCPC², Robinder G. Khemani, MD MsCI², J. Michael Dean, MD MBA¹

¹University of Utah, Salt Lake City, UT; ²Children’s Hospital Los Angeles and University of Southern California Keck School of Medicine, Los Angeles, CA

Introduction
Clinical decision support (CDS) tools link evidence and clinical expertise with patient-specific information; e.g., to make a treatment recommendation that is evidence-based but tailored to the patient state¹. CDS recommendations can only affect clinical care if the recommendations are acceptable to clinicians². Ensuring recommendations are appropriate for the context of care is necessary for clinical acceptance. Knowledge base (KB) maintenance, localization, and adaptation to new use contexts can be an especially challenging aspect of CDS development¹,³,⁴.

Problem
KB maintenance is complex and can require prohibitive amounts of clinical expert and knowledge developer time¹,⁵.

Approach
A CDS tool for ventilator management in pediatric intensive care unit (ICU) ARDS/Acute Lung Injury patients was adapted from a tool used in adult ICUs, based on input from clinical experts in the PALISI and CPCCRN networks. The KB was implemented using production rules; with 2829 rules and 56409 lines of code. We used PICU clinical data to evaluate the extent to which “usual care” actions corresponded with what the protocol would have recommended for the same patient state; using heat maps and other data visualization and analytic approaches to assess correspondence. Emulated CDS tool interactions further explored focused issues with clinical experts. We qualitatively evaluated alignment with the current “parent” ICU protocol and with the emerging guidelines from the Pediatric Acute Lung Injury Consensus Conference (PALICC) group.

FINDINGS: PICU data suggested usual care results in only about 50% conformity to known best practices, with wide variability; so there is substantial opportunity to improve lung protective ventilator practices². The pediatric CDS tool remained qualitatively aligned with the parent protocol and with PALICC guidelines; there were small changes to thresholds and the magnitude of changes but the basic approach was maintained. We identified sets of rules in which

1. Clinical practice already aligns with protocol recommendations. These rules do not need modification.
2. The patient state was seldom observed in the clinical data. These rules were evaluated against literature to ensure that they align with best known evidence and current guidelines.
3. The protocol recommendations do not align with current clinical practices. These rules need to be evaluated by clinical experts. This represented less than 20% of the rules.

Conclusion
KB maintenance and adaptation to new contexts are complex and time consuming tasks but are necessary if users are to accept CDS recommendations. Using clinical data to inform KB maintenance efforts aligns with the vision of the Learning Health System. While knowledge maintenance remains a complex task, clinical data and analytics can usefully contribute to focusing knowledge engineering efforts on rule subsets that might need modification.

References
Understanding Why Providers Override Computerized Medication Alerts in the Inpatient and Outpatient Setting

Michael Swerdloff BS1, Diane L Seger RPh1,2, Mary G Amato, PharmD MPH1,6, Nivethietha Maniam BS1, Olivia Dalleur, MPharm PhD1,4,5, Julie M Fisko BS2, Quoa Liang Her PharmD MS1, Sarah P. Slight MPharm, PhD PGDip1,3, Patrick E. Beeler MD1,9, Tewodros Eguale MD PhD1,6,7, Patricia C Dykes PhD MA RN 1,9, David W Bates, MD MSc1,9

1Brigham and Women's Hospital, Boston, MA, USA; 2Partners Healthcare, Wellesley, Boston, MA, USA; 3School of Medicines, Pharmacy and Health, Durham University, Durham, UK; 4Louvain Drug Research Institute, Universite catholique de Louvain (UCL), Belgium; 5Cliniques universitaires Saint-Luc (UCL), Brussels, Belgium; 6MCPHS University, Boston, MA; 7McGill University, Montreal, Quebec, Canada; 8Massachusetts General Hospital, Boston, MA, USA; 9Harvard Medical School, Boston, MA, USA.

Abstract: Computerized decision support (CDS) alerts are often overridden in the inpatient and outpatient setting. We evaluated both inpatient and outpatient providers focusing on their behavior and practice when overriding these alerts. While providers in the inpatient and outpatient setting were grateful for alerts and felt they contribute to patient safety, inpatient clinicians felt that CDS alerts could be more effective and efficient since they saw their patients more frequently and were likely to experience alert fatigue.

Introduction: Application of computerized decision support (CDS) alert functionality has been variable between inpatient and outpatient providers. We continue to observe a high level of alert overrides. While many overrides are justified clinically, some are not. It is important to reach out to those providers who are not prescribing optimally to understand their reasons for overriding alerts, and whether these reasons differed between the inpatient and ambulatory setting.

Methods: We evaluated all Level 2 alert overrides between January 2009 and December 2011. We then limited our sample to providers who received 20 or more alerts. Of the 2,495 providers eligible for the study, 1,770 inpatient and 725 ambulatory, we targeted those with a high inappropriate alert override rate. Research pharmacists conducted academic detailing sessions tailored to each provider’s override alert profile. A robust and complete analysis of the data was carried out and prevalent concepts related to alert functionality and specific prescribing behavior identified.

Results: We conducted 43 total academic detailing sessions, 34 with outpatient and 9 with inpatient providers. Many inpatient providers felt that alerts reminding them to monitor their patients were extraneous, and ignored these warnings as they were seeing their patients regularly. They recommend the option to disable alerts consistently overridden and ignored. In the outpatient setting, providers found the clinical relevance of the alerts could be improved with provision of recent laboratory values, and the ability to order additional tests directly from the alert.

Conclusion: Insights identified through academic detailing sessions included that alert fatigue existed for warnings deemed irrelevant, and frustration that repetitive alerts could not be disabled. Inpatient providers were more frustrated than outpatient clinicians when prompted to monitor their patients. The alerts were appreciated when the provider first saw the patient, however subsequent alerting was seen as frustrating and time consuming. Inpatient clinicians would like to see repetitive alerts less frequently. By incorporating provider preferences, and considering the increased attention patients receive in a hospital; providers felt that CDS alerts would be less likely to be overridden and more likely to lead to providing more effective, efficient care.

References:

This study was funded by grant #U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ) Dr. Beeler was supported by the Swiss National Science Foundation
Leveraging a Clinical Data Warehouse to improve the quality of data in the French DRG-based system (PMSI)

Emmanuelle Sylvestre, MD, MSc, Christine Riou, MD, Guillaume Bouzillé, MD, Marc Cuggia, MD, PhD

1INSERM, U1099, Rennes, F-35000, France; 2Université de Rennes 1, LTSI, Rennes, F-35000, France; 3CHU Rennes, Centre D’Investigation Clinique, Unité Fouille de Données, Rennes, F-35000, France; 4CHU Rennes, Département D’Information Médicale, Rennes, F-35000, France

Introduction
Over the past few years, clinical data warehouses have been essentially used and developed for research purposes. However, they are a great source of integrated data and can cater to different needs, especially in health economics. Since they are exhaustive and are able to manage heterogeneous data, clinical data warehouses can meet the requirements of the French DRG-based system (PMSI): efficacy (especially regarding the time spent coding) and efficiency (with an emphasis on exhaustiveness). The aim of this study is to develop a methodology to improve coding using drug prescription data from the hospital’s clinical data warehouse.

Methods
A retrospective analysis was performed, using data from the hospital’s clinical data warehouse. All patients hospitalized between October and December 2014 in the otorhinolaryngology ward of Rennes University Hospital, were included. This unit was chosen because surgical records do not necessarily evoke existing chronic diseases when the hospitalization is not related to those. Yet, these diseases are comorbidities often associated with drug prescriptions available in electronic health records but they are rarely mentioned in the surgical record and therefore, cannot be coded. The premise of this study is that it would be possible to infer these comorbidities from drug prescriptions and hence encode them despite their absence in the surgical record.

First, drug prescriptions from the hospital information system were associated with their Marketing Authorizations (MA), using a French drug knowledge database named Thériaque. Thus, each drug was matched with one or several disease(s) and their corresponding ICD 10 codes. This list was then crossed with another one called CMA (Complications et Morbidités Associées), which is a national list of complications and morbidities (and their ICD codes) that significantly increase the cost of a hospital stay. This way, each drug prescription associated with one or several potential comorbidities was identified. Hospitalization data was finally processed: DRG ICD-10 codes from patients with at least one drug prescription were compared to those from the enriched drug database. In case of a missing code, the potential comorbidity was inferred using text mining techniques on electronic medical records (EMR) to find a written proof of the diagnosis in the clinical data warehouse.

Results
39 400 UCD codes were associated with at least one of the 4878 comorbidities from the CMA list. On the 122 hospital stays with at least a drug prescription for this period, 75.4% had at least one drug prescription with no corresponding ICD code. Information on the missing diagnoses with the help of the clinical data warehouse was found in 44.6% of the hospital stays. The three main reasons why information was not found were: the diagnosis wasn’t written but implied, there wasn’t any other proof than the medication on the diagnosis or the prescriptions suggested a symptom that was part of a syndrome which was already coded in the original DRG. The majority of the 58 detected comorbidities (43.1%) were chronic cardiovascular diseases that were not specifically treated (except for the medication) during the hospital stay.

Conclusion
The clinical data warehouse helped retrieve missing DRG codes and therefore improve the exhaustiveness of coded data. Some retrieved codes were part of the CMA list and therefore changed the cost of the hospital stay. This method is generalizable to different hospital areas and need to be tested on a bigger sample.

References
Lost in the Fog: Information Needs in the Care of Patients with Delirium

Teresa Taft1,2, Scott D. Nelson PharmD, MS1,2, Stacey Slager MS1,2, Charlene R. Weir, RN, PhD, FACMI1,2

1Veterans Affairs Medical Center, Salt Lake City, UT; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Problem: Delirium is the most common complication of hospitalized adults over age 651. It occurs in 14-56% of hospitalized patients, depending on the associated conditions2-4. Delirium is associated with increased complications, longer hospital stays, increased rates of institutionalization, and increased mortality5. The annual economic burden in the United States has been estimated as high as $164 billion5-6. Identification of delirium is imperative because it points to a serious health problem requiring very different treatment from conditions with similar appearance7. However, delirium is undiagnosed in 1/2 of cases8. Understanding physician’s decision-making processes in addressing diagnostic dilemmas for delirium is crucial to the design of effective decision support.

Methods: Critical Incident interviews were conducted. Nine internal medicine physicians (5 of whom are experienced geriatricians) at a single location were asked to report a complex case in diagnosing a patient over age 65 with acute mental status changes (AMS). Decision points were identified and decision processes and information needs were elicited by guided questioning. The interviews were audio recorded and transcribed and three authors conducted iterative contextual analysis, using qualitative software.

Results: Preliminary analysis has focused on causes of uncertainty and strategies that physicians use to manage this uncertainty in the delirium workup process. Three main thematic areas associated with uncertainty emerged: (1) baseline mental status data is often unavailable or is hidden, (2) attribution of AMS can be multifactorial & ambiguous, and (3) available information sometimes lacks credibility. Five strategies for managing uncertainty and one “nonuse” also emerged: (1) expending extensive time and effort obtaining records and/or searching in the medical record for the patient's baseline mental status; (2) tracking down people who personally know the patient for direct questioning; (3) ruling out conditions through testing or treatment; (4) making assumptions of dementia when baseline mental status information is not available or is not trusted; and (5) requesting a consult. Despite knowledge of their existence, physicians made no use of delirium decision support tools including: anticholinergic load calculators, lists of medications to avoid in the elderly, or specific delirium screening tools.

Conclusions: There is a need for prior functional/mental status information of healthy patients in the electronic medical record to support delirium workup. Improvements to search mechanisms are needed to improve efficiency in locating mental status information in the medical record. Video and pictures of the patient, such as those typically found on social web sites may be especially useful to physicians in determining a patient’s baseline mental status. Access to medical records from outside health institutions continues to be needed. In addition, the design and accessibility of delirium decision support tools must be addressed to support use within physician workflow.

References
Developing an Enhanced Electronic Referral Management System
Amanda V. Taube¹, Pamela M. Neri, MS², David Kiernan³, Inna Natanel², Harley Ramelson, MD¹,³
¹Clinical and Quality Analysis, Partners HealthCare, Wellesley, MA, United States; ²Partners Healthcare Systems, Inc., Wellesley, MA, United States; ³Internal Medicine, Brigham and Women's Hospital, Boston, MA, United States.

Introduction: While referrals done in an ambulatory environment require the successful completion of multiple independent steps, there is, not infrequently, a breakdown in at least one of the referral management steps that can lead to patient safety and liability issues.¹ The objective of this innovation program is to provide an enhanced electronic referral management module in an ambulatory EHR that would enable practices to meet the requirements for best practice steps for referral management and to observe the effects of electronic referral management on practice workflows.

Methods: An enhanced electronic referral management module was developed in an ambulatory EHR used by a large integrated delivery system in the northeast. The module includes the ability to create, transmit and track referrals throughout the entire referral cycle. Functions in this module include the ability to refer to a practice in addition to an individual specialist, to search a directory of specialists, to maintain a favorite list of specialists and practices, to acknowledge the consult note, to document communications with the patient, to document that the patient has been notified of the consult results and follow-up recommendations and to document the closing of the referral. Advanced system functions include automatically identifying the appointment date by interfacing with scheduling systems, automatically linking the consultant’s note to the referral by connecting with the EHR’s notes module, and using a rule-based approach to identify when a referral is overdue. Enhanced workflow features include the ability to support the staff creating the electronic referral and sending to the provider for authorization or, alternatively, the provider creating the referral and sending to the staff to complete the administrative functions. The module also provides a referral queue for all of the practice’s referrals or a queue focusing only on a particular provider’s referrals. In both of these queues, users can quickly identify what state a particular referral is in and whether it is overdue. The module allows communication with the specialist by providing electronic communication via secure messaging that includes the reason for the referral as well as the ability to include the referring provider’s visit note and the patient’s clinical record (in the form of a C-CDA document). The specialist’s view of the schedule includes an indication that the patient’s visit is a result of the referral with a one-click access to the information entered by the referring physician. Metrics that can be monitored include the percent of referrals that result in a completed specialist visit, the percent of completed specialist visits for which the specialist’s note was electronically acknowledged by the referring provider or staff and 1.2% had documentation that the patient was notified of the consult results and follow-up recommendations.

Results: This study enrolled nine ambulatory primary care practices affiliated with an integrated delivery system in the Northeast. Approximately 80 practice staff members, including physicians, medical assistants, nurses, and administrative staff typically involved in the referral process were recruited across these practices to use the module. After 5 months of use, a total of 2876 referrals were initiated, of which 2127 were closed. Of the closed referrals, 23.8% had a completed specialist visit with a note documented in the EHR, 25.2% of the specialist’s notes were electronically acknowledged by the referring provider or staff and 1.2% had documentation that the patient was notified of the consult results and follow-up recommendations.

Conclusion: A successful referral management program depends upon having the necessary technological tools to properly monitor referrals and identify when a breakdown on the process has occurred. Equally important are the workflow changes that need to be made in the practice in order to monitor the referral queues and intervene when necessary. We have shown that electronic referrals can be successfully created and tracked. Preliminary metrics demonstrate that there are a significant number of referrals that were closed without a specialist note documented in the EHR and that electronic coded documentation of communication with patients is not commonly performed. Additional resources and practice leadership will likely be necessary to address the issues raised by these metrics.

References
Mapping APACHE IV “Reason for Intensive Care Admission” Classification to SNOMED CT
Charat Thongprayoon, MD; Amelia K. Barwise, MB, BCh; Andrew M. Harrison, BS; Brian W. Pickering, MB, BCh; Vitaly Herasevich, MD, PhD
Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC) Mayo Clinic, Rochester, MN

Background: The APACHE IV classification is used to capture admission diagnostic information for calculation of mortality risk in the intensive care unit (ICU). The lack of structured and formal mapping of APACHE IV to standard classification systems, such as SNOMED CT, is a rate-limiting factor for data sharing, aggregation, and classification.1 However, SNOMED CT contains large amounts of information that is not pertinent for most clinical conditions in the ICU. The objective of this study was to build a linked terminology, based on SNOMED CT and APACHE IV, to reduce the complexity of SNOMED CT for linkage purposes, while enabling standardized data registration.

Methods: For the primary reason for ICU admission using the APACHE IV classification system, there are 119 divisions with 422 diagnostic categories. Each reason for ICU admission is classified as non-operative or post-operative, body system or a transplant or trauma-related category, and finally by diagnosis. For mapping purposes, SNOWMED CT Concept Codes were reviewed by a single trained investigator (CT) to identify either a single or post-coordinated SNOWMED Concept match for each diagnostic category of APACHE IV classification using Snow Owl version 2.0.0. Each matched concept-category pair was assessed by categorizing each concept-category pair as a complete match, partial match, or non-match.

Results: Out of 422 diagnostic categories, SNOMED CT provided a complete match with APACHE IV classification for 356 (84%) patients, partial match for 58 (14%), and non-match for 8 (2%) (Table 1).

Table 1: Result of mapping between APACHE IV classification and SNOMED CT Concept Codes

<table>
<thead>
<tr>
<th>System</th>
<th>Number of categories</th>
<th>Complete match (%)</th>
<th>Partial match (%)</th>
<th>Non-match (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>97</td>
<td>77 (79)</td>
<td>15 (15)</td>
<td>5 (5)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>57</td>
<td>51 (89)</td>
<td>6 (11)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Genitourinary</td>
<td>35</td>
<td>26 (74)</td>
<td>9 (26)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Hematology</td>
<td>18</td>
<td>15 (83)</td>
<td>3 (17)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Metabolic</td>
<td>18</td>
<td>16 (89)</td>
<td>2 (11)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Musculoskeletal</td>
<td>19</td>
<td>16 (84)</td>
<td>3 (16)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Neurologic</td>
<td>54</td>
<td>39 (72)</td>
<td>14 (26)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Respiratory</td>
<td>46</td>
<td>41 (89)</td>
<td>5 (11)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Transplant</td>
<td>6</td>
<td>4 (67)</td>
<td>0 (0)</td>
<td>2 (33)</td>
</tr>
<tr>
<td>Trauma</td>
<td>72</td>
<td>71 (99)</td>
<td>1 (1)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Non/post-operative group</strong></td>
<td><strong>Number of categories</strong></td>
<td><strong>Complete match (%)</strong></td>
<td><strong>Partial match (%)</strong></td>
<td><strong>No Match (%)</strong></td>
</tr>
<tr>
<td>Non-operative</td>
<td>212</td>
<td>179 (84)</td>
<td>32 (15)</td>
<td>1 (0.5)</td>
</tr>
<tr>
<td>Post-operative</td>
<td>210</td>
<td>177 (84)</td>
<td>26 (12)</td>
<td>7 (3)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>422</strong></td>
<td><strong>356 (84)</strong></td>
<td><strong>58 (14)</strong></td>
<td><strong>8 (2)</strong></td>
</tr>
</tbody>
</table>

Conclusion: Mapping APACHE IV classification to SNOMED CT Concept Codes can be done in 98% of cases. This mapping result is useful to identify relevant patient subsets using SNOMED CT as a first step in developing the ICU-specific interface terminology.

Reference
Platform for Engaging Everyone Responsibly (PEER) Validation Study Plan

Sheri R. Tibbs, MPH1, Cynthia Kluchar, MS,1 Sharon F. Terry, MA,2 Eric L. Eisenstein, DBA1
1Duke Clinical Research Institute, Durham, NC, USA; 2PXE International, Genetic Alliance, Washington, DC

Introduction
Genetic Alliance (GA) is a network of more than 1000 disease advocacy organizations (DAOs) seeking to advance biomedical research. GA developed the Platform for Engaging Everyone Responsibly (PEER) tool as a web-based application for DAO participants to share clinical information within a trusted environment via access-permission rules. PEER data are aggregated, filtered through privacy directives, and made available to qualified stakeholders.

On January 14, 2015, The U.S. Food and Drug Administration (FDA) approved the Maestro Rechargeable System for certain obese adults. This is the first time patient preference data were used in regulatory decision making.1 PEER is a potential tool for patient preference data collection; however, it requires validation for this use.

Methods
We designed an evaluation study using the UFuRT analysis methodology of Zhang and Butler as adapted to the clinical research data management (CRDM) domain by Nahm and Zhang.2,3 UFuRT is a work-centered design methodology. We used two UFuRT elements (User and Function) to identify PEER user types and their complete functionality set. Each function was standardized as an element in the CRDM domain ontology. We then mapped each standard function to one or more of the PCORI methodology standards to determine which recommended standards should be implemented to ensure scientifically valid patient-centered outcomes research.4

Results
We identified nine PEER systems user types characterized by their clinical research responsibilities and level of technical and research expertise. These included data providers (participants, genetic testing labs / genetic sequencing tool companies and health care providers), application developers (disease advocacy organizations and system administrators), organizations seeking to learn / teach others about a specific medical condition (community organizations, professional societies), as well as interested researchers (standalone and research study sponsors).

We next related each PEER system user type to one of 55 functions from the CRDM ontology developed by Nahm and Zhang.2 We found differences in functional requirements and use intensity between user types. This mapping provided a high-level set of user-derived functional requirements. Lastly, we mapped the previously identified PEER system CRDM functions to elements of the PCORI methodology standards. This second mapping helped us identify the standards that should be adopted to ensure valid results.

Conclusion
We used the UFuRT analysis methodology and PCORI methodology standards to identify PEER system user types, general functional requirements and candidate evaluation methods. While these methods provided an integrated approach, greater detail will be required to create a final evaluation plan and study protocol. Additional research is needed to extend the UFuRT and PCORI methodologies for this use case.

References
A Systems-Based Framework for Informatics Workforce Development to Support Health System Transformation
Herman Tolentino, MD, Laura Franzke, PhD, MPH, Sridhar Papagari Sangareddy, MS
Division of Scientific Education and Professional Development, Centers for Disease Control and Prevention

Introduction

The United States lags behind certain high-income countries in health outcome versus spending per capita [1]. Intersectoral collaborations are needed for participants in the health system to achieve a collective impact for improving health outcomes among both persons and populations. A competent, informatics-savvy, and boundary-spanning workforce is essential to health system transformation to improve population health. To deliver this intended outcome, the authors developed a systems-based framework to support strategic interventions at experiential learning ecosystem leverage points.

Methods

A triaxial framework was developed to represent this experiential learning ecosystem (Figure 1), which has the following components: (1) work and effect settings where collaborative health system transformation and service-learning activities occur; (2) skill development levels [2]; and (3) workforce pipeline stages of professional development. An incremental approach was used to iteratively test different interventions at these leverage points. The continuum of professional development is aided by a semiotic cycle of knowledge generation and application [3], whereby knowledge captured from applied activities in the work or effect settings are organized coherently to further refine and improve interventions.

Results

Multiple interventions have been designed and are at various stages of implementation at leverage points in this experiential learning ecosystem. The poster describes outcomes and effects of ongoing initiatives at these leverage points.

Conclusion

The triaxial framework helped the authors conceptualize and implement innovative interventions at leverage points in an experiential learning ecosystem. The incremental and iterative implementations of these interventions support the strategic delivery of integrated service-learning activities to improve population health.

References
Lack of Unique Healthcare Identifier in Healthcare Information Exchanges: A Field Study
Monica Chiariini Tremblay, Ph.D., Gloria J. Deckard, Ph.D., Peng Zhang

ABSTRACT: This work proposes a two-stage qualitative study on the impact on the lack of a Unique Healthcare Identifier (UHID) on interoperability of patient records and in successful Healthcare Information Exchanges (HIE). We begin by discussing problems introduced by the lack of a UHID: difficulty in correctly matching patient records which results in diminished ability to improve patient outcome through health information exchange. We describe the cases of unmatched records as an instantiation of Type I and Type II errors and examine the barriers to adopt UHID. We empirically study this phenomenon in two field studies conducted in the state of Florida. The research team is part of the evaluation team of the implementation of the HIE in the state of Florida, which gives us access to investigate the statewide HIE transactions; specifically the rates of non-matched and unfound searches. We collect both qualitative data through interviews and metrics related to the State of Florida HIE.

INTRODUCTION: US healthcare system is facing huge challenges. Many scholars, policy makers and practitioners believe that Healthcare Information Exchange (HIE) offers at least partial solution to the problem of interoperability and continuity of care. Despite the great potential benefits that HIE promises, the implementation of HIE is by and large marginally successful and many scholars try to identify barriers that might hinder the development of HIE.

UHID: Among many potential barriers of HIE, some scholars argue that the lack of a unique healthcare identifier (UHID) negatively impacts the development of HIE. If HIE is considered as an IS database infrastructure, it must follow the minimum principles of any good database design. Technically, the HIE database system lacks a primary key, or a unique healthcare identifier (UHID) for all patients. Lacking of a primary key greatly increases the difficulty of identifying a patient’s record, and increases the possibility of erroneous use of patient data in the system. The drawbacks of not having a UHID have been discussed in theoretical or experimental settings by scholars [1], but to date there are few, if any, studies that examine the impact of UHID on HIE in an empirical setting. To gain such insights, we propose a two-wave research project to fulfill such a purpose. By examining the interview data for the evaluation through the three year period, some insights are gained which serves as the foundation for the second round of focused research for UHID.

TYPE I AND TYPE II ERRORS: To exchange healthcare information, such as patients’ medical records, patients are from the lack of understanding of the contribution of UHID to HIE. Our proposed research gain better understanding of UHID and thus facilitate trust of the public and help the development of HIE. The research is comprised of two waves of data analysis. The first wave is based on the analysis of archival data of the Florida HIE evaluation [2] to gain initial understanding of how UHID impacts HIE. The second wave of the data collection of the study, projected to be completed between 2015 and 2016. Specifically, both qualitative data and quantitative data are to be collected to examine the potential UHID impacts on HIE. The goals of data collection is to gain insights of people’s opinions regarding the significance of UHID. PLU’s actual use, or the quantitative data, such as the query time, Type I and Type II errors, and consequences induced by these errors, will be collected and analyzed to quantify the actual loss of lacing UHID or potential gains of when UHID is in place.

CONCLUSION: Despite the great efficiency gains of UHID, the plan of implementation has encountered many oppositions. Many of the oppositions are from the lack of understanding of the contribution of UHID to HIE. Our proposed research gain better understanding of UHID and thus facilitate trust of the public and help the development of HIE.

REFERENCES:
Integrating an Externally Developed Clinical Decision Support (CDS) System with an Existing Electronic Health Record (EHR) System at VA

Samson W. Tu, MS¹,², Kaeli W. Yuen, BS¹, Connie Oshiro, PhD¹,
Susana B. Martins, MD, MS², Ignacio Valdes, MD, MS³, George Welch, MS³,
Paul Heidenreich, MD¹,², Mary K. Goldstein, MD, MS¹,²
¹VA Palo Alto Health Care System, Palo Alto, CA; ²Stanford University, Palo Alto, CA; ³Astronaut Contracting LLC, Houston, TX

Introduction: The VA electronic health record (EHR), Computerized Patient Record System (CPRS), has capability for CDS through its clinical reminders system; however, the built-in clinical reminders are not well designed for encoding complex patient care recommendations. We have previously developed a CDS system, known as ATHENA-CDS, that encodes complex clinical practice guideline recommendations, extracts data from VA’s patient data repository, and, when triggered by events in CPRS, displays in a standalone window guideline-based recommendations for the management of problems such as hypertension and chronic pain. In two experimental projects reported here, we developed mechanisms for a much tighter integration with CPRS and for the delivery of ATHENA-CDS recommendations within CPRS’s clinical reminder system.

Methods: In an initial prototyping project, we worked with Astronaut Contracting LLC to develop Point of Care (POC) software that can invoke ATHENA-CDS for hypertension, export data to it, and retrieve and format ATHENA-CDS recommendations as a dynamically generated CPRS reminder. In a second project, we updated an existing ATHENA-heart failure-knowledge base (ATHENA-HF-KB) to align it with the 2013 American College of Cardiology/American Heart Association guidelines for the management of heart failure and added the KB to the integrated CDS system. The system architecture includes (1) a data-extraction module that extracts clinical data of patients with next day appointments, updates the data and extracts the data of walk-in patients in real time, and converts the data in a format usable by the ATHENA-CDS system, (2) an ATHENA-CDS server that uses the patient data to compute CDS recommendations, (3) a monitoring module that interprets the CDS output to decide whether a patient has ‘reminder due’ or ‘reminder not due,’ and (4) a display module that, for patients with ‘reminder due,’ dynamically generates reminder dialogs and acts on the clinicians’ management decisions.

Results: We expanded and enhanced the knowledge encoded in ATHENA-HF KB to include a greater range of contraindications to our recommended therapies and more nuanced action choices to account for special cases. We implemented the prototype system of integrated heart failure and hypertension CDS within CPRS in a CPRS sandbox provided by the Veterans Health Administration Innovations Office. The prototype integrated system computes recommendations using an external server with encodings of complex guideline recommendations, but allows the VA clinician to interact with patient-specific advisories within VA’s existing reminder system workflow. These externally generated ‘dynamic reminders’ make use of the CPRS capability to include clickable buttons to take actions within CPRS and allow a user to access reference and educational materials through links in the CDS display. We tested the system with a collection of simulated patient encounters.

Discussion: Our work demonstrates how an externally developed CDS service for patients with complex management choices can be integrated in a legacy EHR not designed with such integration in mind. Unlike the earlier integration strategy, which displayed CDS recommendations in a standalone window, the current system embeds the CDS within CPRS’s existing reminder workflow and gives clinicians easy ways to enact guideline-recommended management changes. Aside from the extraction and transformation of data needed for the management of hypertension and heart failure, the architecture can be applied to other clinical areas. A limitation of the experiments reported here is that the CDS display, with the exception of actionable buttons, is confined to formatting available in the underlying EHR, in this case text whose formatting is not as rich as the standalone interface. This work highlights the desirability for an EHR system to facilitate integration with externally-developed CDS services by providing controlled access to patient data and by having mechanisms for interpreting and displaying such computed CDS.

Acknowledgments. The work was supported by a VHA Strategic Innovations Project, and by VA HSR&D grant RRP 12-447 (PI Goldstein). Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs or other affiliated institutions.
Smart Coach: A Problem-Solving Mobile App to Support Weight Loss Management

Bengisu Tulu, PhD1, Emmanuel Agu, PhD1, Stephenie Lemon, PhD2,
Jessica Oleski, MS2, Martinus Evans, MA2, Sherry Pagoto, PhD2
1Worcester Polytechnic Institute, Worcester, MA; 2University of Massachusetts Medical School, Worcester, MA

Introduction

Lifestyle interventions, while effective at reducing weight and diabetes risk, are intensive which makes widespread implementation difficult. Mobile technology may reduce intervention intensity while preserving outcomes by assisting in the delivery of behavioral strategies. However, the range of evidence-based strategies addressed by mobile apps on the current market is narrow.1,2 A key strategy missing across apps in both the market and research is problem solving, an essential component of behavioral weight loss interventions.3 We developed Smart Coach, a mobile app that implements problem solving strategies, to help patients during their weight loss journey by identify solutions to their weight loss problems. In this poster, we report on the usability study results.

Methods

We recruited 29 participants from the local Worcester community for requirements gathering and usability studies. All participants went through telephone screening for initial eligibility assessment followed by in-person screening visit that include consenting procedures, height/weight measurements, and an online survey (measures: demographics, distress tolerance, emotional eating, and emotion regulation). First 12 participants attended a 60-minute “think out loud protocol” where each participant verbalized their thoughts, actions, and feelings while they were using Smart Coach. Next, we conducted 4 focus groups where participants used Smart Coach for 15 minutes and explored its features. A total of 17 participants provided feedback on the Smart Coach user interface, system features, organization of the problem solving process, database of problems and solutions, and ease of use during these focus groups. All participants completed a System Usability Scale (SUS)4 survey at the end of their visits. Our sample was 79% female and 93% Caucasian. The average age of our sample was 49 and the average BMI was 35.

Results

Smart Coach is an avatar-facilitated, idiographic problem solving mobile app that processes information intelligently to help patients identify potential solutions to their weight loss problems. We identified a large set of problems and solutions based on our review of transcripts from 30 weight-loss problem solving session conducted by the investigators. Through an iterative process, we developed Smart Coach utilizing the input from talk through sessions. SUS scores during the development process ranged from 83 to 95. During the focus groups, we observed a wider range for SUS scores (32.5 to 100) with an average score of 79.7 indicating that the user groups were overall satisfied with the features and the content of the app. Most users reported that the app has a pleasant interface, simple and easy to use, and is very intuitive. Users also expressed interest in a social feature and that regularly updated information is important for continued use. However, integrating Smart Coach with their favorite wellness-tracking app was the key factor that would bring users back to using problem solving techniques more often.

Acknowledgement

Research reported in this poster was supported by National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under award number R21DK098556. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References:

Researcher Needs for a Patient-Centered Outcomes Research (PCOR) Data Infrastructure

Petry S. Ubri1, Prashila Dullabh, MD1, Tyne Alexander1, Alison Rein, MS2, Erin Holve, PhD, MPP, MPH2, Abigail Schachter2

1NORC at the University of Chicago, Bethesda, MD; 2AcademyHealth, Washington, D.C.

Research Objective

The Affordable Care Act established the Patient-Centered Outcomes Research Trust Fund (PCORTF), which provides the Office of the Secretary (OS) funding to enable data infrastructure for PCOR. Under the OS PCORTF work managed by the Assistant Secretary for Planning and Evaluation (ASPE), we reached out to research community representatives to gather feedback on key high-level functionalities and standards, policies, services, governance structures, and federal data needed to support a thriving PCOR data infrastructure.

Study Design

From March to April 2014, we conducted nineteen 60-minute semi-structured discussions with researchers, asking them to identify gaps and potential investment opportunities related to functionalities and components needed to enable PCOR data infrastructure. In May 2014, we held an in-person meeting with eleven of the researchers interviewed to identify gaps and priority areas in which further investment in PCOR data infrastructure are needed.

Principal Findings

- Participants called for greater linkages across Medicare and Medicaid data to enable researchers to gain a complete national perspective on patient care, outcomes, policies and services for collecting patient consent.
- Participants agreed that collection of patient-generated health data, including from mobile devices and other sources, is important but highlighted concerns with data provenance and accurate analysis and interpretation of the data.
- Regarding uses of PCOR data infrastructure, researchers’ opinions on the value of some specific investments to increase data capacity differ:
  - Some discouraged further investments in data registries, which they see as workarounds absent standards and interoperability. Others felt registry investments are essential due to data quality and aggregation issues.
  - Some participants emphasized the need for investing in flexible, adaptable common data models to facilitate data aggregation and analysis while others focused on the importance of standards for common data elements.
- Sustainability requires aligning incentives for providers and payers to build, utilize, and maintain PCOR data infrastructure. Participants noted a need for aligning standards and reporting requirements, promoting use of common infrastructure, integrating research processes into healthcare operations, and transitioning to public-private partnerships.
- Participants emphasized the need for feedback loops notifying providers and other data users of quality issues, and metadata standards describing data quality including data completeness, comprehensiveness, and validity. They also encouraged the standardized collection of clinical and claims data.

Conclusions

Researchers’ perspectives on the essential needs, improvements, and mechanisms to support the development and expansion of a PCOR data infrastructure are diverse; yet, there is agreement on the need for investments to address data quality issues, facilitate access to federal and private data assets, enable the collection and integration of patient-provided data, and promote sustainability of data infrastructure through needed standards and policies.

Implications for Policy, Delivery or Practice

By identifying existing gaps and potential opportunities around researcher needs for a PCOR data infrastructure, we highlight potential research and investment opportunities that can improve current processes for data capture, storage, and use of health data for the continuum of activities ranging from healthcare operations through research, including CER and PCOR.
Hospital Participation in Meaningful Use and Rehospitalization of Medicare Beneficiaries

Mark Aaron Unruh, PhD¹, Joshua R. Vest, PhD¹, Hye-Young Jung, PhD¹, and Rainu Kaushal, MD, MPH¹,²,³,⁴

¹Department of Healthcare Policy & Research, Weill Cornell Medical College, New York, NY
²Department of Medicine, Weill Cornell Medical College, New York, NY
³Department of Pediatrics, Weill Cornell Medical College, New York, NY
⁴NewYork-Presbyterian Hospital, New York, NY

Abstract: Nearly 20% of hospitalized Medicare beneficiaries are readmitted within 30 days. Hospitals’ use of electronic health records (EHRs) may reduce readmission rates. Over 90% of non-federal hospitals now possess Meaningful Use certified EHRs. We compared the odds of rehospitalization before and after Meaningful Use for Medicare discharges from Florida hospitals. Compared to the control group, hospital participation in Meaningful Use was associated with 3% lower odds of rehospitalization (Odds Ratio: 0.970, 95% Confidence Interval: 0.949-0.997).

Introduction: The nation has committed nearly $30 billion for provider adoption of interoperable electronic health records (EHRs) through the Meaningful Use program.¹ Over 90% of non-federal hospitals now possess EHRs certified for Meaningful Use.² However, little is known about the effect of hospitals’ participation in the program on the quality and value of care. We examined the impact of hospitals’ participation in Meaningful Use on rehospitalization of Medicare beneficiaries. Rehospitalization among Medicare beneficiaries is frequent, poses risks for patients, and costs the federal program over $17 billion annually.³ Hospital adoption of interoperable EHRs may reduce rates of rehospitalization by improving access to patient information and enhancing the coordination of care. We provide early evidence of the impact of participation in Meaningful Use on 30-day readmissions from hospitals in Florida.

Methods: We linked State Inpatient Databases from the Healthcare Cost and Utilization Project with the American Hospital Association’s Health Information Technology Database and Meaningful Use payment records. The study population included 1,340,387 Medicare fee-for-service (FFS) and Medicare Advantage participants age 65+ discharged from Florida hospitals over the period 2010-2012. Our outcome was all-cause 30-day rehospitalization (yes/no). The determinant of interest was hospital participation in Meaningful Use, identified through payment dates following hospitals’ attestation as meaningful users. Our case group included patients discharged from hospitals that participated in Meaningful Use during the study period. Comparison groups included patients discharged from hospitals that used paper records or EHRs without Meaningful Use participation. The differential timing of hospital participation in Meaningful Use was used to identify the program’s effect on rehospitalization, controlling for patient and hospital characteristics, including EHR use among hospitals that did not participate in Meaningful Use. Estimates were derived from logistic regression models with hospital and year fixed effects and standard errors adjusted for clustering within hospitals. A secondary analysis of Medicare beneficiaries with congestive heart failure (CHF) was conducted to examine patients with a high risk of readmission.

Results: Discharge from a hospital participating in Meaningful Use was associated with 3% lower odds of rehospitalization (Odds Ratio [OR]: 0.970, 95% Confidence Interval [CI]: 0.949-0.997) compared to discharge from a hospital using paper records. In our secondary analysis of CHF patients, hospital participation in Meaningful Use was associated with a greater decrease in the likelihood of readmission (OR: 0.940, 95% CI: 0.893-0.995) relative to the comparison group. In both analyses, patients discharged from hospitals with EHRs that did not participate in Meaningful Use had the same risk of readmission as patients discharged from hospitals that used paper records.

Discussion: Hospital participation in Meaningful Use may reduce the likelihood of rehospitalization for Medicare beneficiaries. Importantly, our study population included both Medicare FFS and Medicare Advantage participants. Our study is among the first to demonstrate the impact of Meaningful Use on the quality and value of care provided to the Medicare population.

References

Automating Personal Health Record Mammography Messages to Improve Mammography Screening Rates

Vaibhav Goyal, MD,1 Peter J Greco, MD,2 and David C Kaelber, MD, PhD, MPH1,2
1Department of Pediatrics, 2Center for Clinical Informatics Research and Education, The MetroHealth System, Case Western Reserve University, Cleveland OH

Abstract: Automated mammogram identification, ordering, and messaging system through standard electronic health record (EHR) and tethered personal health record (PHR) portal tools cant lead to an approximately 50% increase in screening mammograms completed over a 3 month period. EHR tethered PHRs can significantly improve screening mammogram completion rates, but still leave the majority of patients overdue for a screening mammogram without having their mammogram completed within 3 months of messaging.

Introduction: As electronic health records (EHRs) proliferate and their population health tools evolve and as EHR tethered personal health records (PHRs) are implemented and adopted, opportunities will increase to use EHR based population health tools with tethered PHRs to improve patient compliance with recommended preventative health activities. This study uses vendor based (Epic) population health tools and tethered PHR to improve mammography screening rates1,2.

Methods: Using standard commercial EHR functionality (Epic, Epic Corporation, Verona WI), a registry of women was developed who were 1) routinely seen in the last year in a single primary care internal medicine outpatient clinic, 2) due for but had not received a screening mammogram based on standard US Preventatives Services Task Force mammography screening guidelines and 3) subscribed to the EHR tethered PHR (MyChart). Screening mammograms were ordered for these women (so that an appointment could be made against the order) and then batch messages were sent to these women through the PHR. Screening mammography completion rate were compared among these women to women meeting the first two criteria at the same outpatient clinic but not on our PHR during the 3 months, after messaging through the PHR.

Results: 748 women met inclusion criteria for being in the outpatient clinic and for needing a screening mammogram. Table 1 shows the results of the 2 groups during the 3 months after the PHR messaging. Mammography screening completion rates were 15.4% in the PHR group compared to 7.0% in the non-PHR (control) group (p=0.002).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>PHR Group</th>
<th>Non-PHR Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients in Group</td>
<td>421</td>
<td>327</td>
</tr>
<tr>
<td>Mammograms Scheduled</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Mammograms Completed</td>
<td>65</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 1 – Screening Mammography Messaging Study Result

Conclusion: Standard EHR functionality that enables bulk patient identification and test ordering, coupled with bulk messaging to patients through an EHR tethered PHR can significantly improve screening mammogram completion rates, but still leave the majority of patients due for a screening mammogram without having their mammogram complete within 3 months of messaging. The approach has significant applicability to improve patient compliance with numerous preventative health activities.


Development of a concept-based template report editor for Radiological Information System

Carolina Valencia¹; Andrés Arbeláez MD¹, Sebastián Vélez-Zuluaga², Hugo Chvatal¹; Andrés Zapata¹, Lorena Tamayo¹, Ana C. Londoño¹, Jose F Florez-Arango² MD, PhD.
¹. Link Diagnóstico Digital. Medellín, Antioquia, Colombia.
². Grupo Informed, Facultad de Medicina, Universidad de Antioquia, Medellín, Antioquia, Colombia.

Introduction
The Radiological Society of North America (RSNA) has proposed the use of standardized templates for improving radiology-reporting practices. Use of standardized templates represents more specificity, completeness and homogenization of the radiological report’s content, homogenized findings, and is useful for data mining. Person, who creates templates under RSNA standards, requires knowledge of XML/HTML. RSNA templates uses RadLex lexicon (RL) as a reference model. RL is a tool that is not available in Spanish. RL’s terms are domain specific for radiology; there are cases where radiological reports’ intended users are not familiar with the radiological terms. The structure of the RSNA template facilitates translation of reports content into other domains³, as is the case of patients and referring physicians that use different language to express the same concept. RL does not cover all findings and outcomes that radiological reports use. For example RL does not include codes for different classification of arachnoid cysts (Galassi) and grades of appearance of subarachnoid hemorrhage (Fisher).

Methodology
Unified Medical Language System (UMLS) was the base for articulation of controlled vocabularies as LOINC, SNOMED-CT, CIE10 AND HL7. An automated process mapped RL into UMLS using a reference table RL – SNOMED-CT. This mapping allowed some RL concepts to automatically be translated into Spanish. Google’s automated translation service (GATS) served to generate more Spanish concepts. A human expert reviewed GATS and automatic mapping result for correctness. UFuRT framework² was the design and development method used to create a template designer and editing tool (TEDT).

Results
UMLS mapped into 30513 RL concepts. Supervised method provided a total of 24493 terms in Spanish. An intuitive template designer and editor tool is available. The newly created template editor and designer tool (TEDT) uses UMLS+RL ontology as knowledge engine. Templates created in TEDT provide customization of radiology reports while keeping the meaning of findings homogeneous no matter what the language of the intended user is. TEDT allows to new concepts into the Lexicon regardless of the language. Sample reports of Magnetic Resonance Imaging and Computed Tomography modalities served as seeds to build radiological report templates in different subspecialties such as neuroradiology. These templates contain sections and options with predetermined values created by subspecialized radiologists, decreasing the time used writing reports.

Discussion
The radiological and medical knowledge implemented into an interface to create standards-based, best-practice report templates is a step forward compared with the narrative conventional style of radiological reports. This consistently ordered structure helps to avoid errors of transcription (making the report precise, complete, and normalizing its content)³. Moreover the powerful feature of searching reports by pathology, particular finding and any other coded concept in different controlled vocabularies and language, allows data mining, leading to an enormous potential of identifying trends for specific goals in radiology practice, educational and research purposes.

Acknowledgments

References
iDECIDE Smartphone App for Personalized Messages for Nutrition and Fitness Goals
Irene van Woerden, M.S., Lauren Spano, B.S., Pedro Neto, HSD, Danielle Groat, B.S., M. Adela Grando, Ph.D.

Arizona State University, Tempe, Arizona; Federal University of Parana, Curitiba, PR Brazil

Abstract
Current smartphone applications (apps) related to health and fitness allow users to input goals and track progress but they lack personalized feedback to help detect obstacles and suggest alternatives to achieve behavioral changes for a healthier lifestyle. We propose an app, iDECIDE, to help users to identify their ‘persona’ which enables the app to deliver tailored recommendations, provide personalized encouragement and feedback messages and determine goal completion status with the aim to help users achieve their fitness and nutrition goals.

Introduction
It is difficult for individuals who seek to change to a healthy lifestyle to adopt all the necessary changes at once. The process requires that individuals acquire self-management skills by successfully achieving small goals while working towards a larger goal (1). The type of feedback an individual needs for goal achievement depends on their stage in the lifestyle transformation, e.g. athletes may need less encouragement than beginners to improve their physical performance (2). We are developing the iDECIDE app to help users identify their “persona” or stage of readiness to adopt a healthier lifestyle which will allow iDECIDE to provide tailored recommendations and messages to promote appropriate behavioral and lifestyle changes.

Methods
We observed over 9 hours of the Arizona Mayo Clinic “Live It!” program. “Live It!” meetings are facilitated by a registered dietician who delivers information related to weight loss while providing attendee specific information. The emphasis of the program is on educating individuals on how to live a healthier lifestyle.

We also conducted a survey of 28 individuals to acquire insights into the past and current successes and/or failures they experienced with maintaining healthy habits. The survey also asked questions that would help us understand the obstacles individuals face when making and keeping health related goals.

Results
The “Live It!” observations supplemented by the survey responses led us to identify three personas (Figure 1). The aspirant, inspired and athlete personas identify stages in the individual’s health journey and required level of guidance. For instance, the athlete has been working towards a healthier lifestyle for some time; compared with the other personas who may need more guidance and support as they encounter obstacles.

We deployed into the iDECIDE app a questionnaire that assess the persona of an individual (Figure 2). iDECIDE refers to the user’s persona and past encountered obstacles to tailor recommendations and messages, generate challenges and determine goal completion status. For instance, it would suggest a realistic plan when the aspirant-specific decision rules notice that the aspirant user has not completed an exercise goal to run and the obstacle of ‘not enough time to exercise’ is discovered. The app responds with “No worries if you did not find time for a run today, cleaning your house could help you burn some calories too!”

Future Work
iDECIDE will be tested with fictitious case scenarios created for each persona using data from the surveys and Mayo “Live It!” observations. Usability tests will help improve the app’s interface and functionalities.

References
Evaluation of a Statewide Online HIV-HCV-STD Clinical Education Program – Characterization of Healthcare Providers’ Professional Background, Self-Reported Knowledge Increase, and Intention to Change Clinical Practice

Dongwen Wang, PhD; Amneris Luque, MD; Terry Doll; Monica Barbosu MD, PhD; Matthew Bernhardt, MS
University of Rochester Medical Center, Rochester, NY 14642

Introduction
Clinical research in HIV, HCV and other STDs continuously generates new medical knowledge. Timely and effectively disseminating such knowledge to healthcare providers is critical for its translation into clinical practice. Leveraging a history of more than two decades providing in-person training, the New York State (NYS) HIV-HCV-STD Clinical Education Initiative (CEI) (www.ceitraining.org) started its online education program in 2009. Over the past five and half years, CEI has developed 200+ multimedia learning modules, ~100 online CME/CNE courses, a dozen interactive case simulation tools, and various other online resources. These resources are delivered through multiple platforms (websites, mobile apps, email newsletters, and online social networks) to tens of thousands of clinicians (physicians, nurses, etc.). CEI recently released its new student portal as a centralized system to manage online CME/CNE courses and program evaluation. Here we report the initial results to evaluate the CEI online education program, with analyses to characterize healthcare providers’ professional background, their self-reported knowledge increase, and intention to change clinical practice as the impacts of training.

Methods
We selected a study period from Nov. 3, 2014 (when the new CEI student portal was released) to Dec. 31, 2014 (when the quarterly reporting period ended). We queried the CEI student portal on clinicians’ evaluation of online CME/CNE courses with regard to: 1) usefulness/relevance of information; 2) easiness for comprehension; 3) knowledge of the trainer; and 4) appropriateness of the format. We assessed the impact of training through clinicians’ self-reported measures on: 1) intention to use knowledge; 2) knowledge level before and after training; and 3) plan to change clinical practice. For each measure, we further characterized clinicians’ personal (demographics and education level) and professional (occupation, practice years, employment setting and location, caseload, and specific clinical services) background, with Spearman’s correlation test to measure the association.

Results
A total of 335 clinician students completed 12 available online CME/CNE courses in the study period, with very positive evaluations (useful/relevant: 88.66%; easy comprehension: 87.76%; knowledgeable trainer: 91.04%; appropriate format: 84.18%). With regard to impact of training, 83.58% indicated intention to use the learned knowledge, 46.27% had at least 1-level knowledge increase, and 45.78% of those providing direct patient services expressed plan to change practice. Clinicians from the rural areas and with small patient caseload had more favorable experiences of the online training program. The impacts to knowledge use and change of clinical practice had mixed responses, depending on specific courses, clinicians’ years of practice, and particular types of services provided. Clinicians’ demographics, education levels, professions, and employment settings made no differences.

Conclusion
The initial assessment has shown that 1) clinicians are in general very positive about the CEI online education program, and 2) significant proportions of them plan to use the learned knowledge and to change clinical practice. Certain responses may depend on specific courses, clinicians’ experiences, and services provided. Clinicians from the rural areas and with small patient caseload are the groups most benefited from participating in online training. We plan to conduct follow-up studies to assess knowledge retention and to examine the actual impact to patient care.

Acknowledgement: This work is supported by contracts #C023557, #C024882, & #C029086 from NYS Department of Health AIDS Institute, and grant #R24HS022057 from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors. We would like to thank Dr. Xuan Hung Le for his contribution to data collection for this study. We would like to thank NYS and AHRQ program officers Beatrice Aladin, Cheryl Smith, Howard Lavigne, Lyn Stevens, Bruce Agins, and Marian James for their support.
Exploring Healthcare Mobility in the US to Improve Quality of Care: Preliminary Results
Karen H Wang1,2, Constance Carroll1,2, Joseph L Goulet2, Brenda Fenton2, Samah Fodeh1,2, Joseph J Erdos1,2, Marcella Nunez Smith1, Amy C Justice1,2, Cynthia A Brandt1,2
1 Department of Veterans Affairs, VA Connecticut Healthcare System 2Yale University School of Medicine

Abstract, 93 words
Mobile populations, who are often minority and low-income populations, are at risk for poor healthcare outcomes. These populations may experience poor healthcare coordination leading to either delays in or unnecessary care as healthcare systems are not designed for mobility. There are no agreed upon definitions for mobility for healthcare use, and thereby no standard measures and estimates of this type of mobility, or ‘healthcare mobility’. We used Veterans Health Administration national integrated health record systems with clinical data from over 100 facilities to explore healthcare mobility and to identify patterns in healthcare mobility.

Background:
Mobile/migrating populations, who are often minority and low-income populations, are at risk for poor health and healthcare outcomes. These populations may experience poor healthcare coordination leading to either delays in care or unnecessary care as the electronic health records are not personally-owned and as health information systems are not designed for population mobility across distinct border (e.g. regional health information systems). There are no agreed upon definitions for mobility for healthcare use, and therefore there are no standard measures and estimates of this type of mobility, or ‘healthcare mobility’.

Methods:
We used records from one million Veterans randomly selected from the Musculoskeletal Disorder (MSD) cohort, a previously characterized cohort with over five million Veterans who accessed the Veterans Affairs Healthcare System (VAHS) between 1/1/2001-12/31/2013 who have a MSD diagnosis as defined by at least two outpatient visits in 18 months or at least one inpatient visit with an MSD code. Patients were then followed retro- and prospectively from the date of the first MSD diagnosis. Data, extracted from the VAHS Electronic Medical Records, contain demographic and clinical characteristics and healthcare utilization. For this study, we defined ‘healthcare mobility’ as the use of more than one VAHS outpatient facility during the study period. We examined characteristics of mobility based on migration literature, including the frequency of change in outpatient facilities, duration of time per facility, and geographic distance between facilities.

Results: Among this random sample of 1 million Veterans, 14% (n=147,720) had used two or more healthcare facilities. Among the healthcare mobile individuals, there were 82% (n=121,061) with two, 14% (n=20,684) with three, 2.9% (n=4326) with three, and 1.1% (n=1,648) with five or more facility used. Veterans used on average 2.8 facilities. Among the healthcare mobile between two facilities, the median number of total visits alternating between two facilities was 7, and approximately 75% Veterans alternated between two facilities no more than 12 times during the study period. The average duration spent at a facility was 24 months. The figure demonstrates number of uses per year for the sample. To elucidate other characteristics of mobility at more than two facilities we will use more complex analytical tools such as sequential data mining.

Conclusion:
Individuals within the VAHS are healthcare mobile between different outpatient facilities. Care coordination across these facilities is needed to ensure quality healthcare for mobile individuals. Future studies will examine the types of healthcare accessed between the facilities and assess for duplicative services and costs.
Automatically Screening Possible Chemoresistance Genes of Bladder Cancer Drugs

Liwei Wang, PhD1, Jingde Zhu, PhD2, Qian Zhu, PhD3

1Department of Medical Informatics, School of Public Health, Jilin University, Changchun 130021, Jilin, China
2Cancer Epigenetics Program, Anhui Cancer Hospital, Hefei 230031, Anhui, China
3Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD 21055, USA

Abstract
Bladder cancer is highly refractory to the drug therapy due to chemoresistance, which prevents effective chemotherapy [1]. At present the mechanistic understanding of cancer chemoresistance remains elusive [1]. This study aims to predict chemoresistance genes of bladder cancer drugs including Adriamycin, Cisplatin, and Paclitaxel, from a large scale of literature data, SemMedDB. The findings produced from this study would consequently provide novel hypotheses to further investigate bladder cancer chemoresistance mechanisms by biologists and/or oncologists.

Materials and Methods
SemMedDB containing 70 million semantic predications extracted from the PubMed [2], has been explored to identify possible chemoresistance genes. To be specific, a two-step approach has been applied to predict possible chemoresistance genes from SemMedDB, 1) genes associated with “Drug resistance concepts” via specific properties, such as “stimulates” and genes associated with “bladder cancer drugs” via “inhibits” have been extracted automatically and evaluated manually by Dr. Jingde Zhu, 2) one inference rule shown in Figure 1 has been designed for chemoresistance gene prediction.

Preliminary Results
Total distinct 71 genes associated with 15 drug resistance concepts have been identified for the 3 bladder cancer drugs from the SemMedDB. The individual gene list obtained for each bladder cancer drug is shown in Table 1.

Table 1. Sample set of the predicted chemoresistance genes of bladder cancer drugs

<table>
<thead>
<tr>
<th>Bladder cancer drugs</th>
<th>Predicted genes related to chemoresistance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adriamycin</td>
<td>ABCB1, ATP phosphohydrolase, Glycoproteins, Heat shock proteins</td>
</tr>
<tr>
<td>Cisplatin</td>
<td>ABCB1, ATP phosphohydrolase</td>
</tr>
<tr>
<td>Paclitaxel</td>
<td>Cytochrome P450, EPHB2, MAPK1, MAPK8, ABCB1, ABCG2, ADM, APP, ARL4C, ALPK1, Albumins, ATP phosphohydrolase, caspase-3, caspase-9, constitutive androstanone receptor, APP, AHR, ATP phosphohydrolase</td>
</tr>
</tbody>
</table>

Discussion and Conclusion
To our knowledge, this is the first attempt to automatically screening possible chemoresistance genes for bladder cancer drugs from the literature. Further experimental evaluation will be proposed for those predicted genes by Dr. Jingde Zhu’s lab. In addition to the evaluation, more systematic approach by exploring more diverse types of data, including pharmacogenomics, systems biology data and developing more complex inference rules along with the advanced informatics approaches, such as network-based analysis would be used to identify more candidate chemoresistance genes related to bladder cancer drugs or other types of cancer drugs.

Reference
Building and evaluating predictive models for postoperative ileus prior to colorectal surgery
Xuya Wang\textsuperscript{1,2}, M.S., Raul Caso Caso\textsuperscript{2}, B.S., and Yindalon Aphinyanaphongs\textsuperscript{1,2}, M.D., Ph.D.
\textsuperscript{1}Center for Health Informatics and Bioinformatics, NYU Langone Medical Center, New York, NY
\textsuperscript{2}NYU School of Medicine, New York, NY

\textbf{BACKGROUND:} Postoperative ileus (POI) is an abnormal pattern of gastrointestinal motility, frequently occurring after abdominal surgery with an incidence of 3-32\%.\textsuperscript{1} It incurs a significant resource burden on institutions\textsuperscript{2}. In this study, we evaluated a predictive model for POI using variables immediately after the preoperative assessment. Creating a model at this point in time will allow for POI-reducing interventions that include avoidance of mechanical bowel preparation or nasogastric tube placement, performance of minimally invasive surgery when feasible, limiting opioid use, and early postoperative ambulation and feeding.

\textbf{METHODS:} We used the American College of Surgeons' 2012 National Surgical Quality Improvement Program (NSQIP) database, a multicenter (>250 hospitals) database. Of 16,981 colectomy patients, 2,716 (16\%) developed POI. We performed two experiments. For both experiments, we evaluated performance using 5 fold stratified nested cross validation. In the first experiment we evaluated classifier performance using 12 predictors of POI selected from the literature. In the second experiment, we evaluated classifier performance considering all 37 preoperative variables. We compared Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) by examining the areas under the curves (AUC). We compared AUC values with the Wilcoxon rank sum test with adjusted p-values by the Benjamini & Hochberg procedure.

\textbf{RESULTS:} In the first experiment, the mean AUC for the LR classifier was 0.654, 0.653 for LDA, 0.646 for QDA, 0.633 for RF, and 0.550 for SVM. LR was statistically significantly different (Adjusted p<0.05) than RF and SVM. In the first experiment, the top 4 variables in the logistic regression model for one data split were emergency case, dyspnea, gender, and serum albumin. In the second experiment, the mean AUC for the LR classifier was 0.672, 0.671 for LDA, 0.640 for QDA, 0.650 for RF, and 0.595 for SVM. LR was statistically significantly different than QDA, RF, and SVM. In the second experiment, the top 8 variables in the logistic regression model for one data split were estimated probability of morbidity, estimated probability of mortality, dialysis, elective surgery, ventilator dependence, history of congestive heart failure, gender, and prior chemotherapy treatment.

\textbf{CONCLUSION:} The LR model, with a mean AUC of 0.672, was the highest performing classifier with statistically significant discrimination. Model building can be utilized to create a predictive model for POI to be implemented immediately following the preoperative assessment.

The American College of Surgeons NSQIP and the hospitals participating in the ACS NSQIP are the source of the data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.

\textbf{REFERENCES:}


Integrating Analytics and Business Intelligence (BI) into a Health Informatics Curriculum: Pros, Cons and Opportunities

Victoria Wangia-Anderson, PhD1
1University of Cincinnati, Cincinnati, Ohio

Abstract Healthcare is drowning in data as a result of the rapid adoption of electronic health records (EHRs), and health informatics professionals with strong data analytics and BI skills are needed. A survey was administered to gather data about attitudes towards a curriculum that integrates analytics into a Health Informatics curriculum. The results showed positive attitude towards the approach and overwhelming support to maintain the integration and focus on specific areas of analytics.

Introduction Rapid evolution of healthcare data analytics and business intelligence resources has generated information that can be utilized to plan for population health interventions, to reduce healthcare costs, to increase patient centeredness and to improve financial gains. Over the past decade, we have witnessed the active application of computer science and data mining and warehousing tools to data, to progressively convert it to information, knowledge and wisdom. Access to this data and the application of analytics skills can enable healthcare organizations to improve risk analyses, patient outcomes, resource availability, referrals, performance metrics and readmission rates. As electronic information continues to grow exponentially in the healthcare industry due to the rapid adoption of EHRs, there is realization that the data and analytics skills gap already evident in other industries will be a concern for healthcare in the near future. A graduate program in Health Informatics program has begun addressing this need by integrating Analytics and Business Intelligence (BI) in the Health Informatics curriculum. Qualitative and quantitative data was collected from students and faculty to assess their attitudes and perspectives towards the integrated curriculum.

Materials and Methods For this case study, qualitative and quantitative data was collected through a questionnaire. The questionnaire was administered to graduate students, affiliates and faculty (N = 60) teaching in the Health Informatics Graduate program. The questionnaire gathered quantitative data about attitudes and qualitative data about: 1) The benefits of introducing basic and advanced data analytics and BI competencies in the Health Informatics curriculum, 2) Weaknesses, 3) Opportunities, and 4) Priority areas for curriculum focus.

Results The results showed 90% of the respondents agreed that current and future Health Informatics practitioners needed to be equipped with both informatics and analytics skills to balance the focus on technology with data analysis and mining, forecasting and BI. Also, 90% of the respondents had a positive attitude towards the integrated curriculum and the purpose. The respondents all agreed that academic-industry partnership were crucial for staying abreast of the trends in Analytics and BI, and those partnerships needed to be further developed. There was consensus that there is already a data analytics and BI skills gap in healthcare organizations and this would have adverse implication on the organizations, and that health informatics academic programs needed to play a key role in contributing to training the needed workforce. The recommended analytics areas of focus fell in the categories of predictive and descriptive analytics. Areas related to prescriptive analytics were less popular.

Conclusion The integration of Analytics and Business Intelligence in Health Informatics programs is desirable, and professionals in healthcare with health informatics, data analytics and BI skills will be able to make significant needed contributions to the healthcare industry. Industry partners are developing cutting edge health data analytics and BI tools, and academic-industry partnerships are crucial for curriculum to remain relevant. Academic institutions can contribute by training well prepared professionals able to fill the workforce gap.
Pediatric Venous Thrombus Embolus (VTE) Screening Tool: Design, Implementation and Continuous Improvement of a Complex Clinical Decision Support Tool

Emily C. Webber, MD; Indiana University School of Medicine, Pediatrics; Indianapolis, IN
Cara Willey, RN; Indiana University Health Information Services; Indianapolis, IN
Michele Saysana, MD; Indiana University School of Medicine, Pediatrics; Indianapolis, IN
Arash Mahajerin, MD; Children’s Hospital of Orange County; Orange, CA

Background Pediatric venous thrombus emboli (VTE), including both deep venous thrombus (DVT) and pulmonary embolus (PE), is increasing in incidence. The objective of this project was to design a tool integrated into a vendor electronic medical record (EMR) to apply institutional VTE screening guidelines, increase screening, document VTE risk, and provide patient specific clinical decision support (CDS).

EMR tool The VTE tool in the EMR consists of four parts: a screening form triggered by the patient age and admission presented to the physician, clinical decision logic rules calculation of risk, documentation of the risk, and suggestion of patient specific order suggestions based on risk. Recommendations are tailored based on risk factors and contraindications, which were analyzed in a previous publication.

Results The tool was implemented in August 2013 for 12 to 17 year olds, who account for 21% of all inpatient admissions at our institution. Screening rates have been tracked monthly since implementation. During the first 22 months, 2314 patients were screened, with a monthly average of 105 (or 68.3%). Figure 1 illustrates the distribution of risk; the majority of patients screened low risk (91 patients/month average). Moderate risk patients averaged 9.5 patients per month, and high risk patients averaged 2.36 patients per month. Physician orders were analyzed for subgroup of 1318 patients from January to December 2014. During the study period, we identified rates of SCD orders (24.7 average/month) and enoxaparin orders (1.6 average/month). There was an average of 8.75 moderate risk patients and 2.42 high risk patients each month during this time. The assessment found that not all orders were placed using the automated suggestions. There have been no VTE identified in the study population, compared to 3 events in patients who would have qualified for screening in the 12 months prior to the implementation of the VTE tool in the EMR.

Conclusions and challenges Utilizing the integrated EMR tool successfully allowed for screening of our adolescent inpatients and identified moderate and high risk children. Risk stratification is now more consistently documented, and evaluation of order accuracy and efficiency is also being evaluated. Prior to the EMR tool, patients’ VTE risk was not routinely assessed or documented, thus we did not have a reliable baseline of orders correlating to patient risk, however we suspect these higher level of orders for prophylaxis were due to the increased visibility and awareness the EMR tool and associated education. Challenges remain in how to best determine the impact of the EMR tool on ordering practices, since overall VTE prophylaxis orders increased even among patients not screened. Current improvements to the VTE tool include automated presentation at times of clinical status change such as a change in ambulatory status or placement of a central line.

Figure 1. Categories of low, moderate, and high risk for VTE in the screened population.
A method to automatically create titles of clinical notes in electronic medical records

Alan Weiss, MD¹, MBA, Mehdi Rais, MS¹,², Rehal Bojhani MD¹, Dean Sittig, PhD¹,²

¹Memorial Hermann Health System, Houston, TX, USA
²University of Texas Health Science Center at Houston, Houston, TX, USA

ABSTRACT

Background/Objective

There are minimal evidence-based standards to define or benchmark the ease of access to contents of clinical documentation, specifically the clinical note. Outpatient clinical notes help record care, document therapeutic plans, and support a variety of functions that include communication, epidemiology, medical training, medical jurisprudence, and billing. A clinical note title provides a potential glimpse to the issues dealt with on the day of the visit; it can allow providers to quickly identify significant findings, diagnoses and/or results. Without opening and reading the entire note to understand its contents, a title can help providers decide which notes to open and where to focus their attention. In this podium abstract, we look to improve access to the clinical note content utilizing techniques to populate clinical note titles from a database developed from most commonly used billing codes which were translated into clinically relevant abbreviated terminologies, and describe methods to evaluate the impact of these changes. We sought to develop and validate methods to automatically create clinical note titles utilizing clinical and billing system data.

Methods

We identified code sets (ICD9, SNOMED, CPT) and respective descriptions used during the course of visits, and designed and validated a set of rules to create clinical note titles based upon these elements. This research team provided input on a preliminary set of rules to create titles by focusing on the top 95% of codes/descriptions (1200 codes/descriptions in total) used for all patients during this 3 year period. Based on this input, we transformed 2,224,167 clinical note titles from a three year period to assess their performance as compared to the original note title.) This research team is currently assessing the impact of these generated titles by examining the created title against the original title for brevity of final title terms, and accuracy of the title in relation to the note itself.

Results

An initial review by this research team of 200 notes, we were able to create clinically understandable short form through transformation of the billing data in a significant portion of clinical notes. We are currently measuring the feedback from a panel of providers on their perceptions in terms of value versus the alternative title, brevity and accuracy and look forward to sharing those results.

Discussion

We developed a set of rules to create clinical note titles utilizing a data base created from all billing codes used for clinical notes over a three year period within our hospital system which were then converted into an abbreviated medical terminology for insertion into titles. In the absence of missing data such as problem lists, this proved to be problematic when creating clinical note titles, as part of the logic relies upon this; we are exploring options to help resolve problem list gaps. We recognize that there are many other ways to do this (i.e. machine learning techniques) which we explore and invite audience for their thoughts.
Engaging Patients & Families in Contributing Patient-Reported Outcomes to a Pediatric Disease Registry for Comparative Effectiveness Research

Elissa R. Weitzman, ScD, MSc¹ ², Parissa K. Salimian, BA¹, Karen Olson, PhD¹ ², Kenneth Mandl, MD, MPH¹ ², Marc D. Natter, MD¹ ²

¹Boston Children’s Hospital, Boston, MA; ²Harvard Medical School, Boston, MA

Introduction

The voice of the patient is largely missing from large federally funded observational cohort studies for use in medication safety surveillance and comparative effectiveness research. Proven approaches to collecting these data are needed to drive knowledge and advance treatment.

Methods

Participants and parents from 4 Childhood Arthritis and Rheumatology Research Alliance (CARRA) Registry sites were recruited/consented contingent on child diagnosis of juvenile idiopathic arthritis (JIA), lupus, or mixed connective tissue disease and use of a treatment medication. Patient reported outcome (PRO) measures about information needs, adverse drug reactions/events (ADRs)/(AEs), and quality of life were collected using a novel tool that enabled in-clinic reporting via tablet and linkage with clinical registry data.¹ PROs were analyzed to identify participant concerns and data relevant to addressing them. Site IRBs approved the research.

Evaluation Results

186 parents and 88 self-reporting teens have entered data (participation rate 82%). Of parents, 54% have children 2-12 years old, 46% have a teen. Among patients, most have JIA (93%); 78% are female. 38% of parents rated “serious long-term side effects of medication” their top concern, followed by “medication safety” (29%). Other issues (e.g., dose/schedule) were of lesser concern. Aggregate PROs address top concerns and knowledge gaps (Figure 1).

Results: 27% of parents reported a medication related ADR/AE. Of these, 55% (15% in all) reported multiple. Problems were common among methotrexate users: 54% of parents reported child restlessness/irritability, 52% reported stomach ache, nausea, or vomiting.

Conclusion

Findings support acceptability and utility of an informatics-enabled participatory model of engaging a clinical registry cohort in providing health information to answer their prioritized questions. Future work will focus on evaluating feedback of summary information to participants to ascertain value for decision-making, and connect Registry with PRO data to investigate treatment experiences in relation to clinical disease measures. This work is supported by 1R01LM011185.

References

**Transition of Care from an Academic Cancer Center to Community Providers and Survivorship care: Would a Patient Care Team Portal Help?**

Philip E. Whalen MA, Marina Oktapodas, and Carmelo Gaudioso MD PhD Roswell Park Cancer Institute Buffalo, NY

**Abstract**

In the transition of care from an academic cancer center to community providers and survivorship, cancer patients’ continuity and quality of care are susceptible to the vagaries of communication, care coordination, and the relationship between providers. The aims of the present study were to a) characterize patients and provider’s experience with transition of care (e.g., needs, preferences, issues, and concerns); b) discern their attitudes and openness toward a patient care team portal designed to facilitate communication and coordination of transitional and survivorship care; and c) identify the essential functional requirements of the proposed patient care team portal. This research hopes to better future interactions between patients, academic cancer centers, and community providers through a patient care team portal designed to facilitate the sharing of information and knowledge in such a way that the transition of care and survivorship care processes are improved.

**Introduction and Background**

Following the HITECH act of 2009, many healthcare institutes began the construction and use of electronic medical records (EMR). Because of this, adoption of portal technologies, which have the potential to fulfill meaningful use requirements of the act, has become widespread with twenty-six billion dollars of a total government allotted thirty-three billion dollars of incentive money pledged as of September 2014 (CMS Data). While a patient portal has been shown to serve as an effective patient-provider communication tool, it requires optimization guided by well characterized users’ preferences and concerns to effectively facilitate transition of care and survivorship care.

**Methodology**

Qualitative survey data were collected during single sessions of focus groups following guided and open discussions examining the potential benefits and issues of using a patient care team portal during transitional and survivorship care. Surveys and discussions were conducted with three separate groups: 42 cancer patients, 32 community physicians, and 7 academic cancer center physicians. Themes were identified using a methodology presented in E.H. Bradley et al., 2007, utilizing conceptual codes and inductive reasoning.

**Results**

A few main themes and concerns emerged during group discussion of the use of patient care team portals in support of survivorship and transition of care: a) a need for better communication between all three cohorts; b) a lack of medical information sharing and care coordination between all three cohorts; and c) a lack of trust between the community physicians and academic cancer center physicians regarding the return of the patient to community care. Themes A and B were endorsed by all cohorts while theme C was primarily endorsed by the two physician cohorts. As such, provider-patient and inter-provider communication, sharing of medical records and disease specific educational information, and guidance were suggested as features of the portal to alleviate these concerns and provide a more coordinated care team as well as an overall improvement in transition of care and survivorship.

**Conclusion**

Recognizing the limitation of the small number of academic physicians, the present study demonstrates the strong favor of stakeholders in the development of a patient-provider portal. Main themes and issues were identified and solutions were developed. Using these qualitative data, we plan to implement and test in real time a module of an existing portal as a central information hub to promote coordination of care and effective communication between all stakeholders.
Increasing Patient Enrollment in Clinical Trials Using a Web Based Recruitment Application

Nancy Whipple, BSN, MS, MM1, Inna Natanel, BA1, Harley Ramelson, MD, MPH1,2,3
1Partners HealthCare Inc, Wellesley MA; 2Harvard Medical School, Boston, MA; 3Brigham and Women’s Hospital, Boston, MA

One study found that in the United States, 34% of the trials recruited less than 75% of their planned sample.1 The primary barriers include: patients’ lack of awareness of relevant trials (40% of adults report that they do not understand the idea of a clinical trial), patients’ fear of being a test subject; concerns about cost and insurance coverage, and general fear of the unknown. Traditional methods of recruitment can be both costly and difficult to anticipate or control.

At Partners HealthCare Systems (PHS) there are over one million outpatient visits and over one hundred thousand inpatient admissions each year. Yet only a small percentage of patients are involved in research studies and clinical trials. In the spring of 2015 PHS will launch a new secure website to assist in the recruitment of patients into clinical trials. Health On-Line for Patient Enrichment (HOPE) is a tool that is designed to provide patients with a view of current clinical trials. It will simplify the process of communication between the patient and the trial administrator and will potentially increase the number of patients enrolled in clinical trials. HOPE will be accessible via a link that is prominently shown on the home page of the PHS Patient Portal: Patient Gateway. Because Patient Gateway is supported by the patient’s clinicians, patients have a high degree of confidence in the content of the site and a feeling that the site is a credible source of information. Patients will receive a secure mail message from the HOPE Team inviting them to explore the site.

The HOPE site will directly address patient concerns and fears. The site will describe the numerous benefits of participating in clinical trials. The patient will be guided through the process of expressing interest in a clinical trial. The site will also provide detailed information about the trial, so that the patient can be more informed about the specific trial. This Information includes clinical eligibility, patient responsibilities (time commitment and activities), duration of trial, and compensation.

Patient experience goals:

- To provide a sense of community, focused, disease-specific, health areas patients may join are created. Community membership provides patients with targeted educational material and events (e.g. webinars, teleconferences), as well as a list of all active trials and the patient’s relationship to them.
- To keep patients engaged, periodic messages are sent that are of particular interest to them. Recipients are based on the communities patients have joined and trials they have expressed interest in. In addition, educational materials, web conferences/dial-in informational sessions, and news items are updated frequently.
- To entice patients to periodically return to the HOPE site, a list of timely, newsworthy items related specifically to HOPE and to clinical trials in general is maintained.

Associated with the HOPE site is an administration tool to maintain the content of the site. The tool also provides the ability to update the status of the patient in relation to a trial (e.g. enrolled, ineligible, and dropped out). We will also be able to report on patient activity on the site. Standard reports will provide usage of specific functionality including measuring the interest in particular health areas, educational materials and events. The analysis will include determining the effectiveness of the various components of the site and determining both the rates of trial enrollment and the rate of retention of enrolled patients. It is our hypothesis that trial enrollment will increase and effective recruitment will lead to long term patient retention.


1735
Augmenting Psychiatric Care: A Participatory Mobile Framework

Kandace M. Whiting, DNP, RN-BC, James Rosser, LCSW, Kenneth Wells, MD, MPH, Armen Arevian, MD, PhD
UCLA Resnick Neuropsychiatric Hospital, Los Angeles

Introduction

Mobile interventions are increasingly used to augment care and facilitate home-based communication, real-time monitoring, and just-in-time interventions. As more mobile health interventions are implemented, there is increasing evidence that a “one size fits all” approach has significant limitations. Community-partnered participatory techniques (CPPR) improve patient engagement and clinical outcomes; however, they are resource intensive and costly to implement in given the iterative nature of intervention technology development. The Chorus Participatory Mobile Framework (CPMF) is designed to address these challenges by allowing individuals to create mobile health interventions quickly at low cost using a simple web interface.

Aim

The UCLA Neuropsychiatric Hospital Partial Hospitalization Programs (PHP) provide psychiatric services for adults and children with acute mental illness on an outpatient basis. Though patients receive extensive support during program hours, support provided after program hours is limited. Consistently, patient satisfaction metrics highlight after hours support as an area for program improvement. To address this gap, we developed patient-tailored mobile interventions such as SMS reminders and mobile web applications to provide patients additional support after hours.

Methods

Using the CPMF, we partnered with providers and patients to develop patient-tailored mobile interventions such as SMS reminders and mobile web application. To do so, we invited PHP providers and patients to in-person workshops to develop initial mobile intervention tools. These workshops involved an iterative cycle of four steps: 1) identification of barriers and opportunities for improvement, 2) generation of messaging content, 3) creating working prototypes of the mobile tool, and 4) testing mobile tools in real-time. Developed tools are customizable and intended to be specifically tailored to individual patient treatment plans, communication preferences, and experiences. After initial prototypes were developed, we offered PHP patients the opportunity to receive additional support after hours by partnering with providers to develop individualized interventions using the CPMF in real-time. Therefore, patients were able to directly create the mobile interventions that they would then use.

Conclusions

The CPMF provides a novel method for providers and patients to co-create mobile health interventions in real-time designed to augment outpatient psychiatric services. This method combines the best practices of community-partnered participatory research principles along with patient-centered design to ensure that health interventions are relevant to individual patient needs. Importantly, CPMF meets a key methods gap by offering an opportunity to implement scalable methods to engage patients, families, and providers through personalized and adaptable mobile health interventions.

References

A Centralized Data Collection and Management Tool in the VA: REDCap

Erika Whittier, MPH,1 Denise M. Hynes, MPH, PhD, RN,1,2 Peter Watson, PhD,3,4
1VA Information Resource Center, Edward Hines, Jr. VA Hospital, Hines, IL;2College of Medicine and School of Public Health, University of Illinois at Chicago, Chicago, IL; 3University of Colorado School of Medicine, Denver, CO; 4Denver VA Medical Center, Denver, CO;

Introduction:
The Department of Veterans Affairs (VA), Office of Research and Development (ORD), supports health research at more than 150 VA facilities nationwide.1 This research has significantly contributed to advancements in health care for Veterans and created innovations that advance health care for the nation. Collaborations that bring together clinicians, program leaders, and researchers within VA facilities, across multiple VA sites and across service areas are critical to this success, and finding efficient ways to bridge research coordination across these areas is essential. REDCap, Research Electronic Data Capture, a web-based application software that allows users to build, share and manage databases and surveys quickly and securely,2 is an instrumental part of bridging these efforts.

Methods:
The VA Information Resource Center (VIReC) partnered with the VA Informatics and Computing Infrastructure (VINCI) to manage the implementation of REDCap within the VA. REDCap is used widely by groups (including National Institutes of Health and numerous academic affiliates of VA Medical Centers for research purposes). Many VA investigators have experience with REDCap through their academic affiliations, thus the community of current REDCap users within the VA is large.

Sought a data management tool that would allow integration across multiple project teams within the VA that: 1) Resides in a centralized location, 2) Provides a secure way of collecting and sharing data in a web-based environment, 3) Provides a secure tool to develop surveys and databases from conception to production on the Web without additional software requirements for researchers.

The implementation began with phased roll-outs across the VA. REDCap was first tested with five groups of testers: three experienced with REDCap, two groups with no experience. Groups were given details instructions and each asked to complete project specific tasks in REDCap. The next phase of Beta testing surveyed 30 beta test users to examine utility of the system, utility of education materials and users’ experience.

Conclusions:
VA REDCap is currently available to researchers with Institutional Review Board Approved research protocols and Quality Enhancement Research Initiative Investigators. Users are able to set up databases, and collect and manage data which is stored on a centralized secure server. Users are also able to generate surveys that can be administered through the VA’s intranet. There are currently 257 projects and 592 users. Examples of current projects include a database for managing data collected from a pilot intervention to improve care coordination between VA and non-VA providers using the Blue Button feature of MyHealthE Vet and a cross-sectional study designed to examine suicide risk assessment practices of VHA Suicide Prevention Coordinators.

References
Routine Collection of Patient-Reported Data in Electronic Form in Clinical Settings: An Analysis of Available Technologies

Adam B. Wilcox, PhD; Susan Tew, BSN; Justin Poll, PhD
Intermountain Healthcare, Salt Lake City, UT

Introduction

Studies of the relative contribution of different determinants to health estimate medical care as a minor contributor, while social circumstances and behaviors contributing the majority of the difference in health outcomes. These factors often must be assessed from patients directly. However, if they are not recorded in electronic form, the data are not available for decision support, longitudinal monitoring or retrospective analysis. Collecting patient-reported data in electronic form is thus critical for transforming the healthcare system.

Currently, most collection of patient-reported data is done with paper forms, but must be transcribed to electronic form. The high cost of transcription makes this approach rarely sustainable. Patient portals can collect information from patients directly in electronic form, but are rarely used and do not support clinical workflows. Two additional methods for patient data collection, that support clinical workflows but do not require manual transcription are scanned forms and tablet-based data collection. Here we describe our analysis comparing scanned forms and tablet computers for collecting patient-reported data.

Methods

Three experts in data collection applications, patient-facing technology, and data collection methods performed this analysis. First, we identified a list of important characteristics important for data collection technologies. For each characteristic, we then rated the two technologies on a 3-point scale (positive, neutral, negative) in terms of how well the characteristic was supported by the technology. Characteristics were then grouped by functional categories, and the technologies were compared across the different categories.

Results

Thirteen characteristics among 4 functional categories were compared. Overall support was equal between scanning and tablets for categories of institutional cost, security, and patient acceptance. However, for functionality, tablets were better than scanned forms for 5 of 6 characteristics. These results suggested a strong advantage for tablets over scanned forms.

Characteristics where tablets provided better support were data security, patient matching, workflow integration, data timeliness, branching logic, extensibility, and patient preference. Scanned forms had lower theft risk, had better disaster recovery support, and lower patient-perceived security risk. The technologies were perceived as equivalent in overall equipment costs, infection control risk, and training costs.

Discussion

Our analysis concluded that tablets were better than scanned forms for collecting patient-reported data in electronic form. The differences were primarily due to advantages in functionality, where support for branching logic and extensibility were important advantages for tablets. Others have studied different technologies for data collection including tablets and scanned forms, however, these studies presumed use of the technology by research data collectors, rather than by patients. Many of the categories used were consistent among the studies, which supports the overall conclusion of the analysis. Future work should focus on demonstrating the use of tablets for patient-reported data collection, and analysis of the adoption and consistency of use.

References

Using Informatics Tools to Standardize the Request, Adjudication and Monitoring of Non Formulary Agents at a VA Facility
Lisa Winterbottom MD, MPH, Judith McConnachie MPH, Thomas Brenk MPA, Jianji Yang PhD, Brian M Wilcox, Pharm.D
VA Portland Healthcare System, Portland, Oregon

Introduction
VA’s electronic medical record system does not have a standardized platform to order non-formulary medications. Facilities developed their own methods to review, approve and order these medications, usually with consult submission to approving pharmacists. The request, adjudication process and monitoring vary by site. When the requests are approved, patients are placed on medications that are most likely high cost and/or have significant safety concerns, requiring close oversight and monitoring. Review of this care requires manual chart reviews that are labor intensive and may not be consistently done for all patients on non-formulary medications. This is a well-recognized patient safety and cost issue in all facilities and authors submitted a proposal and were awarded with a VA Innovation Grant. The grant allows us to collaborate with VA Pharmacy Benefits Management (PBM) to re-engineer and standardize the non-formulary medication request/approval and patient monitoring processes, leveraging capabilities of clinical informatics tools to efficiently support real-time population monitoring. The program will consist of three components: 1) development of standardized tools to support non-formulary medication ordering, review, and approval for use; 2) development of a registry to support population level and efficient patient-level clinical monitoring; and 3) human intervention by a dedicated pharmacist to use the developed tools for appropriate review, approval, and monitoring of patients. The program will be piloted at the VA Portland Health Care System (VAPORHCS) with a robust evaluation to determine impact. After initial piloting, the program will be extended to additional sites in the region with consideration for national use.

Proposed Methods
Program development: A clinical informatics team at VAPORHCS will partner with PBM to develop the standardized Consult template prototype in VA’s Computerized Patient Record System (CPRS) to collect clinically relevant data based on the medication’s “criteria for use” at the time medication request. This information may include the medication name, indication, planned duration, last labs etc. as well as specialized fields of information for a specific drug or drug class. We will partner with VISN 21 on the use of their developed adjudication process involving a uniform note using several searchable data fields. A web-based non-formulary medication monitoring registry will be developed to support two aspects of monitoring. Firstly, it will help track the status of all requests from submission to approval/denial. Secondly and importantly, the registry will support close monitoring of the patients on any non-formulary medications at a population-level as well as a detailed patient-specific level for safety reviews by assembling and displaying clinically relevant information at a glance. Built-in registry logic will also highlight possible concern areas to alert reviewers of priority reviews needed. A dedicated pharmacist will both review requests for the designated non-formulary medications as well as use the registry for close monitoring and timely intervention as needed for patients using the designated medications. Interventions may include stopping medications, titrating doses, directing patient to provider review, etc.

Pilot and Evaluation: The program will be tested in the VAPORHCS with selected medications. Evaluation will be performed to assess improvements in safety monitoring and early intervention when indicated (e.g. doses adjusted, labs ordered, etc); staff satisfaction with process for monitoring non-formulary medications; and financial impacts. Rollout to additional test facilities and nationwide: If the evaluation indicates positive impact, the program is expected to roll out to additional test facilities and then nationwide.

Expected Impact and Current Status
We expect Veterans will benefit due to the improved patient safety from active monitoring. Providers and pharmacists will experience a more efficient monitoring process. Facilities may see a reduction in medication cost due to the standardized request process and better management. The architecture of the informatics tools (CPRS consult and registry) proposed (Figure 1). Input on drug candidates solicited; chart reviews being performed to assess identification of data points. Partnerships with external and internal stakeholders are established. Tools to be tailored to drugs selected with goals of scalability, feasibility, sustainability.
A BI Tool to Monitor the Intervention Efficiency of Antibiotic Therapy in Leukemia Patients
Cai Wu, MS
The University of Texas MD Anderson Cancer Center, Houston, TX

Abstract
A BI tool with Hyperion Reporting Studio was developed to monitor the efficiency of the intervention process for the MDACC antimicrobial stewardship team on daptomycin, linezolid, vancomycin usage in leukemia patients. Daily defined dose (DDD) per 1000 patient days, days of therapy (DOT) per 1000 patient days, Cost per 1000 patient days, mean duration per 1000 patient days, and patient length of stay (LOS) were the measurement indexes. The SPC control chart of the each index was generated to measure the efficiency of the intervention on the antibiotics.

Introduction
The use of broad-spectrum antibiotics is a well-known risk factor for infection in bacterial and fungal in leukemia patients. Increasing of duration of patient exposure to antimicrobials increase the likelihood of colonization with resistant organisms. The University of Texas MD Anderson Cancer Center (MDACC) antimicrobial stewardship aims to optimize antimicrobial use in hospitalized cancer patients by involving a continuous effort of healthcare providers. Its goals are to improve outcomes, minimize therapy cost, and reduce the resistance of antimicrobial use. In order to monitor the efficiency of the intervention process and analyze the duration of antibiotic therapy in leukemia patients, a business intelligence (BI) tool with Hyperion Reporting Studio was developed for Antimicrobial stewardship team.

Measurement Indexes
Based on WHO Anatomical Therapeutic Chemical (ATC) Classification System Defined Daily Dose (DDD), DDD per 1000 patient days, days of therapy (DOT) per 1000 patient days, cost per 1000 patient days, mean duration per 1000 patient days, and patient length of stay (LOS) were used for the measurement of the efficiency of the intervention process. The selected antibiotics were daptomycin, linezolid, vancomycin. Patients were in hospitalized adult leukemia patients (Age >= 18 at the antibiotic dispensing date).

Monitoring Process
The data sources were pharmacy Centricity database and institutional enterprise data warehouse. Hyperion Interactive Reporting Studio 11 was used to query the weekly patient antibiotics usage and LOS from both data sources before and after intervention. The antibiotic therapy duration was based on the antibiotic dispensing starting date of the patient visit, and LOS was based on the admission starting date of the patient visit.

With selected time period, all antibiotic measurement indexes were calculated, the mean, upper control limit (UCL) and lower control limit (LCL) of each drug were generated and charted for the intervention statistical process control (SPC) chart (Fig. 1); Mean and median duration therapy and patient LOS chart was created for easy viewing (Fig. 2).
Visualizing Clinical Workflow using Time and Motion Data

Danny T.Y. Wu, MS¹, Nikolas Smart, MS², Sang-Jung Han, MS¹, Maria Majeed, MS¹, Jin Han, MS¹, Suinan Li, MS¹, Fan Zhang, MS¹, Kai Zheng, PhD¹, ²
¹School of Information; ²School of Public Health; University of Michigan, Ann Arbor, MI

Abstract

Understanding clinical workflow is crucial to improving quality, safety, and efficiency of patient care delivery. A common source of data for studying clinical workflow is through time and motion studies, which however generate multi-dimensional datasets that are very difficult to analyze. To address this challenge, we used three visualization techniques to visualize time and motion data to allow for easy identification of important workflow patterns. These visualizations help us better analyze physical movements, relationships between locations and activities, and the magnitude of change in workflow.

Introduction

Studying clinical workflow to detect deficiencies and optimize processes has great potential to improve quality, safety, and efficiency of patient care delivery. Clinical workflow is commonly characterized using data collected via time and motion observation, which has been a gold standard for quantitative workflow studies.¹ Because time and motion data contain many dimensions, such as locations, timestamps, activities, and roles, challenges remain in effectively analyzing such data to identify important workflow patterns. To address this challenge, we used three visualization techniques to explore potentially interesting clinical workflow patterns from a time and motion dataset that we collected in an ambulatory primary care practice. These visualizations were designed to allow for easier and more effective human inspection of patterns related to physical movements, relationships between locations and activities, and the magnitude of change in workflow.

Method

The dataset contained 68 observation sessions whereby independent human observers shadowed clinicians to collect time and motion data. Each session lasted 3.2 hours on average. The clinician participants included four providers, two nurses, one assistant and one receptionist. Their activities were categorized into three levels: task, category, and theme. Three visualization techniques were used to support the analysis of the data.

Results

Figures 1–3 show the three visualizations. Figure 1 is a path graph illustrating clinicians’ physical movements in the clinic. Figure 2 is a sequence sunburst chart illustrating the relationships between locations and activities. Figure 3 is a horizontal multi-bar chart illustrating the magnitude of workflow change between two calendar months in which the observations took place. We are currently using these visualizations to guide our analysis of important workflow patterns in the study clinic.

Conclusion

We implemented three visualization techniques to identify important workflow patterns in a time and motion dataset. The analysis is ongoing and the findings will be validated using other sources of workflow data collected from the same clinic.

Reference

Use of an Electronic Health Record as a Research Tool: Frequency of Exposure to Targeted Medical Conditions and Health Care Providers’ Clinical Proficiency

Tim Wysocki, PhD, Center for Healthcare Delivery Science, Nemours Children’s Clinic, Jacksonville FL
James Crutchfield, PhD, Mission Systems and Training, Lockheed Martin Corporation, Orlando, FL
Maria Carmen G. Diaz, MD, Division of Emergency Medicine, Nemours/Alfred I. duPont Hospital for Children, Wilmington, DE
James P. Franciosi, MD, MSCE, Division of Gastroenterology, Hepatology and Nutrition, Nemours Children’s Hospital, Orlando, FL

INTRODUCTION: Pediatric medical care frequently requires complex decision making and nuanced judgments that tend to deteriorate over time if those skills are not used. Frequency of exposure to specific health care problems is known to be an important influence on maintenance or decay of health care providers’ clinical proficiency in managing those problems. The electronic health record (EHR) could be used to quantify frequency of exposure and to select clinical problems for study that provide opportunities to understand 1.) Whether health care providers’ clinical proficiency in managing a specific health condition tends to decay with increasing duration since the prior exposure to that condition; and 2.) Which characteristics of the work, the worker and the workplace affect the magnitude of decay in clinical proficiency. This study sought to demonstrate utility of the EHR in identifying and forecasting specific clinical skill domains in care for specific targeted medical conditions that may yield the greatest benefits in terms of quality and safety of care.

METHODS: This research was conducted at Nemours Children’s Health System in the Delaware Valley and Florida. The data reported here were obtained at the health care provider level in a completely de-identified, aggregated manner from Nemours EHR Data Warehouse for the most recent 2-year period. The criteria for selection of targeted clinical problems for further study were: high frequency of care encounters; varied types of providers; adequacy of EHR; variable frequency of exposure to that problem within providers; presence of evidence-based practice guidelines; variability in health care providers’ performance; controversy among experts about optimal management; and interest among clinical divisions to facilitate this research. Based on those criteria, the research team selected (from an initial list of 100 candidates) nine clinical problems for further study, including obesity, influenza vaccination, gastroesophageal reflux disease (GERD), concussion, supracondylar fracture, idiopathic scoliosis, headache, encopresis-constipation and type 1 diabetes mellitus. Health care providers, recognized within Nemours as having special expertise on each condition were recruited to serve as subject matter experts (SMEs) for a series of task analysis interviews, often supplemented with published practice guidelines, designed to define optimal clinical management of each problem and to specify EHR data elements that could quantify expression. The team developed a Δ-t statistic, time since the provider’s prior clinical encounter with the targeted clinical problem, as an index of frequency of exposure. The team also collected certain worker (e.g. health care provider demographics, time since prior board certification, EHR proficiency) and workplace (e.g. on-call schedule, clinic appointment density) variables from sources other than the EHR itself.

RESULTS: The nine clinical problems were evaluated across the Nemours healthcare system, using aggregated data from 2013. Encounter frequency ranged from 1566 cases of encopresis-constipation to 220,774 encounters for influenza vaccination. Mean Δ-t ranged from 1.72 days, for obesity, to 30.79 days for influenza vaccination. Based on the relatively low number of cases, encopresis-constipation appeared with the next highest value with a mean Δ-t of 23.25 days. Maximum Δ-t ranged from 285, for obesity to 497 for headaches. Obesity, influenza vaccination and GERD demonstrated a maximum Δ-t of somewhat less than a year, but the rest indicated greater than 400 days. Except for influenza vaccination, the distribution of Δ-t for each clinical condition fit a Gamma distribution (P < 0.001), indicating that the appearance of patients for each of these conditions follows a Poisson distribution. The distribution for influenza vaccination differed due to seasonality of administration, but still generally fit the distribution. Number of different clinicians seeing these conditions averaged 250, with Type 1 diabetes having the smallest number of different professionals, at 114.

DISCUSSION: This project demonstrates the utility of the EHR as a research tool in studies of quality and safety of care and specifically frequency of exposure of health care providers to specific health care problems. We believe that the use of the EHR to characterize frequency of exposure of health care providers to specific health care problems will allow health care organizations to be better prepared to identify intervention loci that can lead to improved quality and safety of care in a cost efficient manner. Subsequent steps in our research include building a multivariate model to enable prediction of clinical skill decay/maintenance and to test interventions focused on preventing decay of clinical knowledge and skills.
PheWAS Network Analysis and Visualization

Yaomin Xu, Todd Edwards, Lisa Bastarache, Rebecca Jerome, Shilin Zho, Eric S Torstenson, Wei-Qi Wei, Jana K Shirey-Rice, Erica Bowton, Yu Shyr, Jill M. Pulley, Joshua C. Denny
Vanderbilt University, Nashville, TN, USA

Abstract
Understanding the genetic determinants of complex diseases is essential in personalized genomic medicine. In this work, we will demonstrate an exploratory data analysis framework, integrating networks and statistical modeling techniques, to build a set of graphic representation and analysis strategies depicting complex associations between genetic architecture and disease risks. We will illustrate with a case study of our approach applied on a dataset from Vanderbilt EMR-linked biobank with 30,000 subjects and 1500 disease codes.

Introduction and background
In genetic association studies of complex diseases, we expect many genes affect each disease and each gene affects many diseases, and together they form a networked, many-to-many relationship that is challenging to characterize or make sensible inferences. Dense phenotypic data from a DNA biobank linked to Electronic Medical Records provide unique opportunities to unravel the complex disease genetic associations(1). Network modeling provides an effective technique to model complex data relationship as graphs and facilitate exploratory data analyses and data visualization(2). We aim to utilize the network modeling integrating genetic architecture and disease associations to characterize the complex 1) disease similarities and 2) gene-gene interactions from a phenome-wide association study (PheWAS).

Method
Assuming that disease risk varies among different sub-populations determined by genetic variants, we first characterize the gene-based genetic architecture using a coherent genetic network. Second, we calculate the risk profiles over the estimated genetic architecture. Third, we quantify the disease similarities and genetic interactions based on the correlation of the calculated risk profiles. Fourth, we model the disease similarity or gene-gene interaction using weighted networks. Finally, we analyze the network topology to identify the interesting disease genetic relationship such as coherent disease clusters of a given gene and genetic modules of a given disease. We model this against the PheWAS performed for variants predicted to confer risk to several autoimmune disease.

Results
Our approach assembles phenome-wide disease genetic associations into 1) a gene-centric disease networks that characterizes the relationship of diseases based on their risk patterns over the genetic networks, 2) a disease centric genetic networks that characterizes the gene-gene interactions based on risk profile among individuals, and 3) a gene-disease bipartite networks that characterizes many-to-many relationships of multiple disease with multiple genes. Specifically, in our association, the visualization allows data-up assembly and testing of haplotypes associated with disease.

Conclusion
In summary, we presented a novel network-based approach to facilitate PheWAS data visualization and analysis and help better understand complex disease genetic associations.

Simplified Spectrographic Display for Bedside Electrographic Seizure Detection in the ICU

Peter Z. Yan, M.D.1, Zachary Grinspan, M.D., M.S.2
1 Department of Neurology, New York-Presbyterian Hospital - Weill Cornell Medical Center, New York, NY
2 Center for Healthcare Informatics and Policy, Department of Healthcare Policy and Research - Division of Health Policy and Economy, Department of Pediatrics - Division of Child Neurology, Weill Cornell Medical College, New York, NY

INTRODUCTION: A seizure is a temporary dysfunction of the brain involving excessive synchronous neuronal discharge. A seizure lasting more than 5 minutes (status epilepticus) is life threatening. In ICUs, seizures are often non-convulsive (up to 1/3 of patients in neurologic ICUs), detectable only on EEG. Early treatment is key, as more than 80% of patients respond to treatment if it is initiated within 30 minutes, but less than 40% respond if treatment is delayed for two hours or longer. Mortality (33%-57%) can double if seizures are untreated after 24 hours.1 As such, continuous EEG (cEEG) is commonly used to identify seizures in medical or neurologic ICUs. However, this produces hours of EEG records to be read by a sub-specialized neurologist, who gives report to the ICU physician usually 1-3 times daily, a workflow which may delay treatment.2

A simplified seizure visualization display for bedside ICU physician could potentially reduce this delay. Current EEG visualization systems are primarily designed for the neurophysiologist; these systems are complicated and poorly suited for physicians without sub-specialized training.

METHODS: We applied two spectrographic variations to EEG data of 186 seizures from 24 pediatric patients: a median power spectrogram (MPS; a novel method) and a hemispheric asymmetry spectrogram (HAS; adapted from published research and commercial software.) MPS is less sensitive to artifacts and serves as a generalized seizure detector (figure 1). HAS serves as a simplified focal seizure detector (figure 2). In combination, we expect most seizures will appear in at least one of the spectrograms. Power spectograms were produced with short time Fourier transforms (hamming window, window length = 2s, FFT length = 212, 50% overlap). The median power of each frequency bin per minute was calculated for the median power spectrogram. The sum of the power on the left minus the right was calculated for the hemispheric asymmetry spectrogram. EEG data was acquired from the CHB-MIT scalp EEG database.3 All signal processing was performed with MATLAB. Spectrograms were qualitatively reviewed for harmonic resonance, or vertical “ripples”, indicative of electrographic seizures as well as other obvious changes from background.

RESULTS: One non-subspecialist (PY) identified 88% (162 of 186) of seizures in an initial non-blinded review. This is compared with published reports of 79-83% seizure detection involving traditional color density spectral array (CDSA) by non-subspecialists or certified neurophysiologists.4,5 Qualitative review of the EEG waveforms suggests that majority of missed seizures (14 of 186) consisted of slow delta seizures on the EEG.

CONCLUSIONS: Preliminarily, the combination of a median power and hemispheric asymmetry spectrogram detect seizures more frequently compared to CDSA in published reports. This raises the possibility that the MPS+HAS display will be easier to use and more sensitive in seizure detection when compared to current methods.

NEXT STEPS: A robust evaluation with multiple blinded readers is in progress, which will include estimates of sensitivity, specificity, and inter-interpreter variability. A new graphical user interface is also in development to facilitate review of the spectrograms. The eventual implementation will be integration into an ICU dashboard, such that electrographic, physiologic, and pharmacologic data will be displayed side by side in real time to assist the ICU physician with clinical decisions.

Figure 1. (A) EEG tracing with generalized seizure (B) Median Power Spectrogram (C) Hemispheric Asymmetry Power Spectrogram. Lines marks seizure duration.

Figure 2. Different patient (A) EEG tracing with focal seizure (B) Median Power Spectrogram. (C) Hemispheric Asymmetry Power Spectrogram. Lines marks seizure duration.

References
GeoHealth Informatics: Applying Geographic Information Science (GIS) to Support Heart Failure Self-Care

Nan Ye, M.S., Daniel Johnson, Ph.D., Richard J. Holden, Ph.D.
IUPUI, Indianapolis, IN

Research question
Patients with heart failure experience challenges performing daily self-care activities such as exercise, taking medications, monitoring for symptoms, diet, and less frequent activities such as attending appointments and managing medication supply. This study’s research question is, how can geographic information science and systems (GIS) facilitate heart failure self-care and mitigate self-care barriers?

Methods
Based on recent studies of heart failure self-care and secondary analyses of data from 61 older adults with heart failure, we identified common geospatial factors shaping self-care. We then identified opportunities for GIS to facilitate self-care through a combination of searches of: 1) existing software and devices with GIS functionalities and 2) literature on GIS applications to chronic illness self-care.

Results
Diet management was influenced by location in several ways, including the proximity of fast food restaurants and distance from sources for healthy foods. Being located in one’s home versus eating out affected diet, although this depended on patients’ circumstances such as the presence of a family member at home. Physical conditions such as weather and steep terrain, lack of nearby exercise facilities, and inadequate transportation affected patients’ ability to exercise or attend cardiac rehabilitation sessions. Distance from the pharmacy, pharmacy and clinic hours, and comfort driving were examples of geospatial factors influencing medication management. Various other geospatial barriers were identified; some (e.g., transportation) affected multiple self-care behaviors. Several existing GIS software applications and tools in the market can potentially assist heart failure self-care, e.g.:

- **On-demand transportation options** (e.g., Uber) could overcome distances and lack of personal vehicle
- **Wearable/personal devices** with GPS and other sensors can detect problems, connect patients to clinicians and caregivers, and provide feedback on entering a ‘geo-fenced’ area (e.g., vibrate near low-walkability street).
- **Food system maps** (e.g., [http://mdfoodsystemmap.org/map/](http://mdfoodsystemmap.org/map/)) can show food deserts and guide patients to healthy food resources or allow patients to recommend or rate local restaurants based on sodium content.
- **Geospatial data collected from a panel of patients** can inform clinicians of the challenges and resources that patients, individually and in groups, may encounter in their places of work, residence, and daily activity.

Conclusion
Barriers to and resources for heart failure self-care vary by physical-spatial, social-cultural, and organizational context. GIS technology can help patients better understand, organize, and access resources to make informed self-care decisions and overcome self-care barriers. This poster will demonstrate opportunities for GIS to improve health and disease management by leveraging GIS tools to help patients with chronic disease bridge specific self-care gaps. Challenges to be discussed include device cost, tool usability issues, patient privacy, uncertainty about patient behavior and self-care adherence from GIS data, the need for resources specific to older patients and patients with heart failure, and issues related to caregiver support. We argue for developing what we call GeoHealth Informatics, an area that crosses health informatics with health geography to support better health and disease management.

References
Intelligence in Usability Survey Research (iUSuR): an Online Usability Question Bank for Usability Survey Research

Po-Yin Yen, RN, PhD¹, Nima Esmaili Mokaram, BS², Gina Torelli, BS³, Alissa Schultz, RN, MA, MS¹,⁴, Xiaonan Ji, MS¹,²
¹Dept of Biomedical Informatics, ²Dept of Computer Science & Engineering, ³Dept of Industrial Engineering, ⁴College of Nursing, The Ohio State University, Columbus, OH

Abstract
Because of low cost and easy administration, usability survey research is commonly used to evaluate health information technology (HIT) for general user satisfaction or technology acceptance. However, support and resources for usability survey research are limited. We created an online usability question bank to help researchers search and select appropriate instruments for their HIT usability survey research.

Introduction
Usability survey research uses survey instruments (aka questionnaires, scales, or surveys) to evaluate HIT. Numerous survey instruments have been developed and validated. However, these survey instruments are often changed or manipulated.¹ It is common for researchers to create new and/or modify existing instruments, because of missing or dissimilar interpretation of usability concepts from existing surveys, or lack of awareness of current, validated instruments. We collected validated questionnaires for technology evaluation and developed an online usability question bank, Intelligence in Usability Survey Research (iUSuR)² to assist researchers conducting usability survey research.

Methods
Currently, iUSuR contains 246 questionnaires consisting of 6970 individual questions. We established a database using MySQL, storing information of entity, referential, and domain integrity constraints, and set up a database driven web application using PHP and CodeIgniter Framework. The application uses object-oriented and component-based structure with model-view-controller (MVC) architecture pattern to organize and present the website. We stored three types of data from each survey instrument: 1) full citation - including author(s), year, title, journal, and cited count; 2) usability concepts measured, validation methods and types, full questions, and answer types; 3) advanced features to have individual questions tagged with usability taxonomy, to enhance the categorization. These advanced features also enable searching across the survey instruments, to identify physical features, personal interactions, task performance, organizational roles, personal characteristics and intention to use.³

Results
iUSuR allows researchers to browse and select existing survey instruments, or aggregate questions from multiple instruments as needed (Figure 1). When used as a collaborative tool, iUSuR’s analytic system tracks usage data on all survey instruments. Each time a user selects and/or modifies a survey questionnaire or question, it automatically returns anonymous usage data to the question bank. Users can base their decision on the usage data to justify selection of an original or modified question or, secondary modification of the question for their individual study needs.

Conclusion
iUSuR is an inventory of usability survey instruments as well as a collaborative site for researchers. Future plans include the implementation of an expert system to provide usability concept recommendations; thus assisting low cost, efficient, and effective HIT usability survey research.

References
2. Intelligence in Usability Survey Research (iUSuR): http://uniter-bp01.bmi.osumc.edu/iusr/beta/
Annotating Recommendation Sentences in Radiology Reports

Meliha Yetisgen, PhD1,2, Prescott Klassen, MS2, Lucas H. McCarthy, MD3,1, Tom H. Payne, MD4, Martin L. Gunn, MB ChB5

1Biomedical and Health Informatics, 2Department of Linguistics, 3Department of Neurology, 4Information Technology Services, 5Department of Radiology, University of Washington, Seattle, WA

Abstract

Communication of follow-up recommendations when abnormalities are identified on imaging studies is prone to error. The absence of an automated system to identify and track radiology recommendations is an important barrier to ensuring timely follow-up of patients especially with non-acute incidental findings on imaging studies. We are in the process of building a natural language processing system to identify clinically important follow-up recommendations in free-text radiology reports so that the reports can be flagged and separate workflow processes can be initiated to improve the consistency and quality of care delivery. To accomplish this, we are creating a multi-institutional radiology report corpus annotated for clinically important recommendation information.

Dataset

We created a large-scale radiology report corpus containing 745,058 reports from three different institutions including Harborview Medical Center, University of Washington Medical Center, and Seattle Cancer Care Alliance. The corpus covers a range of imaging modalities, including radiographs, computed tomography, ultrasound, and magnetic resonance imaging.

Annotation Process

We define recommendation as a statement made by the radiologist in a given radiology report to advise the referring clinician to further evaluate an imaging finding by either other tests or further imaging. When we analyzed our corpus, we observed that there are various different types of recommendation mentions in radiology reports. Based on this analysis, we grouped the recommendations under the following four categories; (1) Non-contingent clinically important recommendation (e.g., CT chest is recommended to further evaluate the lung mass.), (2) Contingent clinically important recommendation (e.g., If this breast lesion has not been previously evaluated, recommend referral to the breast service for further evaluation), (3) Clinically important recommendation likely reported (e.g., L distal radius fracture x 1 week, please also follow-up to rule out scaphoid fracture compared with last week's x-rays), (4) Clinically unimportant recommendation (e.g., Consider an MRI of the forearm if diagnostic certainty is desired).

Using the BRAT rapid annotation tool, two clinical experts annotate the boundaries of recommendation sentences and assign each annotated sentence to one of the four categories. Because manual annotation is a time-consuming and labor intensive process, we can only annotate a small portion of our large radiology report corpus. To make the process faster, we use a semi-automatic annotation approach where we train an initial statistical classifier on a small set of labeled examples we created for preliminary work. By manipulating the configuration of the classifier, we build a very high recall (90.18%) but low precision model (35.07%). Details of this seed dataset and classification approach can be found in [1]. We run this model on a randomly selected set of reports from our un-annotated corpus and present the annotators only reports that include recommendation sentences identified by the model. The annotators read the complete reports to correct system-generated annotations or add new annotations.

Conclusion

Our two expert annotators annotated 50 reports for recommendation sentences. The interrater agreement on the first set of 50 reports is Kappa 0.43 and an F1-score of 0.59. One of the main reasons for these low agreement scores, is the subtle differences in the perception of importance of recommendation sentences by the annotators as they assign categories. To resolve the disagreements, we scheduled multiple meetings. The annotators completed the annotation of 567 radiology reports using updated guidelines based on inter-annotator agreement stage. They highlighted 265 sentences as category 1, 90 sentences as category 2, 222 sentences as category 3, and 160 sentences as category 4.

Acknowledgements

This work was supported by NIH - NIBIB Grant Number 1R21EB016872.

References

Anchor time extraction for building timeline from Korean clinical narratives

Wangjin Yi¹, Seongkeun Park, MD. Ph.D² and Jinwook Choi, MD. Ph.D²,*
¹Interdisciplinary program in Bioengineering, Graduate School, Seoul National University; ²Department of Biomedical Engineering, College of Medicine, Seoul National University

Introduction: To build a timeline extraction and visualization system¹ for Korean clinical narratives this research focuses on recognizing the anchor time as the first part of whole system. A clinical narrative contains valuable information on a patient’s progress of a disease. However, it is hard to efficiently extract information on Korean narratives since free word order of the language makes its expressions complex to be processed.

Anchor Time Extraction using Conditional Random Fields: An anchor time presents one or more events with a group on a certain point in a timeline, and provides clues for following the progression of a patient’s history. We define the two types of anchor times as points (date, time) and intervals written as time expressions. Since they are morphologically similar to lab values, to distinguish them precisely, we formulate the task of extracting anchor time as a sequence labeling. One of the most popular sequencing model, Conditional Random Fields (CRF)², is used to enhance contextual label structure. The best performance in a similar task (i2b2 2012) is F1 of 0.91³ by using CRF.

Tokenization by letter types: In most cases, two or more distinct tokens are in a singleton term in our corpus (e.g., number + Korean). Since the features in the CRF are based on tokens, we additionally tokenize singleton terms by letter types to achieve better generalization.

Feature extraction: Features used in this research are grouped as three types; 1) morphological features (Korean morphemes of the current token), 2) dictionary features (Temporal noun lists, modifiers around temporal expressions), 3) contextual features (token context within window size 3, the occurrence of white spaces)

Experimental Results: We randomly selected 600 discharge summaries of patients, who admitted for any chief complaints and discharged in January 2013, from the electronic health record system of the Seoul National University Hospital. The collection was divided to three distinct sets and we tested the CRF extraction system on each set. Based on exact matching, the performance was evaluated by Precision (P), Recall (R), and F1 (F) with 10-fold cross validation for each set. The performance on the set is 0.897 (P), 0.831 (R), and 0.862 (F) (averaged value). For performance validation, we additionally used a different kind of narratives and same kind but written at different time; 1) 200 admission notes of patients who discharged in January 2013, 2) 200 discharge summaries of patients discharged in 2003. Especially, clinical narratives written in 2003 were created right after the adoption of the EHR in the hospital, and contained writing errors making them noisy for processing. The performance on the admission notes in 2013 is 0.922 (P), 0.856 (R), and 0.888 (F). And the performance on the discharge summaries in 2003 is 0.946 (P), 0.893 (R), 0.919 (F). They are achieved when we use 2nd order CRF.

To evaluate the effectiveness of the tokenization method, we compared the performances of the extraction systems trained on each text set tokenized by white space and letter types for both set of discharge summaries. We confirmed overall improvement in F1 when the extra tokenization was applied (+0.196, +0.129 for 2013 and 2003 respectively).

Limitation: This research performed on a small set of clinical narratives created in a single institution. However, to overcome this limitation we gathered various kinds of narratives and tested on multiple settings.

Conclusion: To recognize anchor time from Korean clinical notes a CRF extraction system was used and it is highly effective based on a text tokenized by letter types. And the model is more effective when the test sets are noisy.

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

References
Annotation of Disease Characteristics for Cancer Liver Stage Prediction

Wen-wai Yim¹, Sharon W. Kwan, MD², Guy Johnson, MD, PharmD²
Meliha Yetisgen, PhD¹,³
¹Biomedical and Health Informatics, ²Department of Radiology, ³Department of Linguistics, University of Washington, Seattle, WA

Abstract
Liver cancer is one of the leading cancer-related causes of death worldwide that occurs across different lifestyle and genetic environments. Despite availability of many new treatments, deaths due to liver cancer have increased in the United States even as those of other cancers have decreased and more than 80% of patients in all stages do not survive past 5 years. To facilitate evidence-based research in liver cancer, we (a) define a set of domain-related characteristics to annotate, (b) construct a workflow to support annotations at several levels, and (c) evaluate the consistency of our annotation scheme through a double annotation experiment. In this abstract, we describe our annotation process, as well as our inter-annotator agreement for our task.

Introduction
Liver cancer is a deadly disease with only a 17% 5-year survival rate across all stages, which occurs over different lifestyle and genetic environments. In the United States, liver cancer incidences are projected to increase with larger contributions of obesity-related cases. Because the liver is a vital organ involved in multiple roles, liver cancer often occurs along with other liver problems such as cirrhosis. Therefore, there are many competing risks of mortality that affect treatment options. To determine the best treatments given certain patient characteristics, a large diverse dataset must be collected and reviewed. However, the process of manual chart review is time-consuming and expensive. In order to build an automatic system of liver cancer stage extraction, we annotated a cohort of patients from the University of Washington Medical Center for liver cancer disease characteristics. To do this, our clinical experts first identified 14 parameters to annotate and created annotation guidelines for marking up clinical documents at both a text-span and patient level. After constructing our annotation workflow, we tested inter-annotator agreement on a sample of 20 patients.

Annotation
Our domain experts identified a set of 11 entity categories along with specific values each can take on for text-span annotation: (1) ascites—accumulation of fluid in peritoneal cavity (e.g., “no significant ascites”) with values (None, Mild, Moderate-Severe), (2) Child-Pugh—a measurement of liver cirrhosis (e.g., “Child’s B”) with values (A,B,C), (3) ECOG (Eastern Cooperative Oncology Group) Performance Status—a scaled measure of general well-being (e.g., “ECOG 0”) with values (0,1,2,>=3), (4) Extrahepatic Invasion—spread of cancer outside of the liver (e.g., “No evidence of extrahepatic extension”) with values (No, Yes), and several other entities. In some cases, non-explicit text evidence was highlighted (e.g. Ascites: None, “no free fluid in abdomen”). Patients not meeting cohort criteria were also identified for removal during annotation. Annotated report types included a variety of clinical notes (e.g. admit, interventional radiology, surgery), as well as CT or MRI abdomen radiology reports. Each liver cancer entity also had a corresponding patient level annotation to resolve possible conflicting information represented in different text-spans annotated for the same entity. In addition, BCLC (Barcelona Clinic Liver Cancer), CLIP (Cancer of the Liver Italian Program), and AJCC (American Joint Committee on Cancer) liver cancer stages were annotated at the patient level. Annotations for text-span and patient level labeling were performed using brat, a web-based annotation software, and an in-house python user interface software, respectively. Two annotators, both interventional radiology attending physicians, annotated text-spans for the 11 entities in the clinical notes of 20 patients. Afterwards, patient level annotations were annotated by the consensus of the two annotators. Of the 20 patients 3 were excluded due to irrelevant diagnosis. Inter-rater agreement showed an entity micro f1-measure of 0.910 at the patient level, 0.854 at the document level and 0.729 at the partial text-span level.

Future Work
We are in the process of annotating a 236-patient corpus to build an automatic liver cancer stage predictor. However, our methods may be useful for other types of cancer prediction, clinical cohort selection, and medical concept recognitions systems.

Acknowledgements
University of Washington Medical Center / Harborview Medical Center, Department of Radiology, Health Services Research (HSR) Section awards, Institute of Translational Health Sciences UL1TR00042
Predicting Autonomy for Physical Activity using Data Mining Techniques
Sunmoo Yoon, RN, PhD1, Niurka Suero-Tejeda, MS, MA1, Suzanne Bakken, RN, PhD1,2
1School of Nursing, 2Dept. of Biomedical Informatics, Columbia University, NYC, NY

Abstract
We applied data mining techniques to a community-based behavioral dataset to develop prediction models for gaining insights for a social media-based intervention for an urban underserved population. Environmental and modifiable personal factors influenced both positive intrinsic motivation ‘willing to make time for physical activity’ and negative intrinsic motivation ‘hard to make time for physical activity’. Data mining methods were useful to build physical activity prediction models prior to designing self-management interventions.

Introduction
Lack of physical activity poses a serious health threat to racial and ethnic minorities. In a previous study, we found that autonomy (‘willingness to make time’) influenced achievement of physical activity goals. The Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project aims to build an infrastructure to understand health behaviors to improve the health among an urban underserved population. Twenty-eight percent of those who completed the WICER survey reported use of social media. Consequently, social media has the potential to be an interventional platform for promoting physical activity. To examine this potential, this study aimed to develop prediction models for autonomy for physical activity by applying data mining techniques.

Methods
We extracted data on the sample participated behavioral survey (n=5,653) from the WICER research data warehouse using the REDCap. 122 behavior variables were iteratively selected from 945 variables by a domain expert (SY). 106 variables were further selected after applying M99.0 algorithm to remove useless attributes. Six unique variables were selected using CFS attribute evaluator1 to discover variables that were strongly related to physical activity. We applied a classification algorithm (J48) to iteratively generate models with random 10-fold validation. We selected final models based on predictive ability and clinical meaningfulness of variables.

Results
Environment factors (having place for exercise, and availability of fresh fruits and vegetables in the neighborhood) and personal modifiable factors (computer time and perceived weight) were the main predictors of willing to make time for physical activity (accuracy 74%, AUC=.72). Those with having place for exercise were more likely to willing to make time for physical activity. Among those with having no place for exercise, having fresh fruits and vegetables in their neighbor, perceived weight as normal and computer use ≤5hours/day was associated with increased likelihood of willing to make time for physical activity (Figure 1). Factors influencing negative motivation included cost barrier for exercise, self-confidence of work ability (score 1-5), and the social behavior of attending meetings (accuracy 65%, AUC=.66%). Participants with cost barrier and self-confidence score of work ability ≤3 were more likely to have negative motivation (Figure 1).

Conclusion
Data mining methods were useful for domain experts to build prediction models for autonomy of physical activity.

Acknowledgement: This study was funded by WICER4U R01 HS022961 (PI: Bakken)

References
Using TURF Framework to Improve EHR-CPOE Medication Dosing in Renal Impaired Elderly

Rafeek A. Yusuf, MBBS¹, Mandana Salimi, BS¹, Min Zhu, PhD, MD¹
¹The University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX

Introduction

Older adults (age>65 years) will account for 20% of the United States population by 2030. In older adults, renal impairment is evident due to decreased glomerular filtration rate, renal tubular function, and creatinine clearance¹. In 2011, the United States spent $49.3B on End Stage Renal Disease (ESRD) programs. Some studies show the positive impact of Computerized Physician Order entry (CPOE) in reducing mortality rate. Studies have revealed the necessity of adjusting the dosage of medications in the presence of renal insufficiency. If it is not adjusted, patients' lives may be put at risk. CPOE supports electronic ordering of medications. However, CPOE with poor user interface may result in technology-induced errors and pose risks to patients. As such, it is important to evaluate the usability of CPOE user interfaces, identify potential issues, and based on well-established framework, design new interfaces that address identified gaps.

Methods

First, the authors evaluated an existing EHR-CPOE module used for prescribing for older adults with renal impairment. Evaluation for usability issues was performed both electronically and manually using: (1) Usability heuristics, (2) User testing, and (3) System Usability Scale (SUS). Second, the authors used TURF framework² to analyze and redesign the existing EHR-CPOE interface, then developed a prototype with a graphical software. Finally, both existing and redesigned interfaces were evaluated for efficiency using the Keystroke Level Model-Goals, Operators, Methods, and Selection (KLM-GOMS) method.

Results

The existing EHR revealed poor usability and learnability. 11 violations of 6 types were noted; average severity rating of these problems was 3 (i.e. major violation); and SUS score was 10. For example, there were too many objects and related objects were not placed properly on the existing interface. The redesigned interface was less cluttered and related objects placed in close proximity. 25 steps were involved using existing versus 7 for redesigned EHR-CPOE module. Redesigned solution took 16.92 seconds versus 61.77 seconds for existing module to prescribe. Frequency distribution of KLM operators utilized in the existing versus redesigned solution of the EHR-CPOE module are shown in the bar graph below (see figure 1).

Conclusion

Use of evidence-based heuristics for evaluation of optimal EHR user interface design requires testing under realistic conditions. Usability testing can be used to demonstrate association between different usability problems and specific medical errors.

![Figure 1. Bar graph comparing KLM-GOMS operators between existing and redesigned EHR-CPOE.](image_url)

References

A Preliminary Study on EHR-Associated Extra Workload Among Physicians

Jing Zhang, MS1,2, Kellie Avery1, Yunan Chen, PhD2,3, Shazia Ashfaq, MD, MBA2, Steven Rick, BS2, Kai Zheng, PhD4, Nadir Weibel, PhD1, Harry S. Hochheiser, PhD6, Charlene Weir, PhD7, Kristin Bell, MD5, Mark T. Gabuzda, MD4, Neil Farber, MD1, Braj Pandey, MD1,2, Alan Calvitti, PhD2, Lin Liu, PhD1, Richard Street, PhD9,10,11, Zia Agha, MD1,2,4

1University of California, San Diego, San Diego, CA; 2Veterans Medical Research Foundation, VASDHS, USA; 3University of California, Irvine, Irvine, CA; 4West Health Institute, La Jolla, CA; 5University of Michigan, Ann Arbor, Michigan; 6University of Pittsburgh, Pittsburgh, PA; 7University of Utah, Salt Lake City, Utah; 8VA San Diego Healthcare System, San Diego; 9Texas A&M University, College Station, TX; 10Baylor College of Medicine, Houston, TX; 11Houston VA Center for Innovations in Quality, Effectiveness and Safety

Introduction

While EHR systems can improve the efficiency and quality of clinical practice, the use of EHRs is commonly associated with increased workload such as prolonged documentation time and new types of documentation tasks1, resulting in end user frustration. Nonetheless, few studies have examined what accounts for the extra workload. This work aims to understand the sources of the extra workload brought upon by the use of EHRs. Doing so, we hope to provide sociotechnical solutions that can alleviate the increasing workload faced by healthcare providers.

Materials and Methods

Semi-structured interviews were conducted with 14 physicians at the VA San Diego Healthcare system. Interview questions were aimed at understanding how physicians use EHRs before, during and after clinical encounters. We used affinity diagram, a common method in qualitative studies, to analyze the interview data2.

Results and Discussion

The analysis shows that the use of EHRs led to extra work for physicians. Most physicians complained that they now spend more time during and outside of visits to catch up on EHR-associated tasks. We categorize the sources of the extra workload into the following categories: 1) Poor usability. Physicians have to perform many clicks to enter orders, and also have to manually enter information for medication lists, which can be time consuming. 2) New tasks enabled by EHRs, such as how there are an overwhelming amount of clinical reminders built in the systems that require physicians to review and approve. 3) Changed work responsibilities. The convenience of EHRs resulted in new work responsibilities for physicians, such as how primary care physicians now have to serve as cosigners for a variety of tasks in the EHRs. 4) Administrative tasks. EHR systems have various quality assurance and administrative functions built in, which require extensive documentations by the physicians. More analysis is needed to evaluate the new tasks imposed on healthcare providers, and to carefully redesign the work responsibility, the functionality and usability of the EHR systems. As part of a larger study, we plan to recruit more participants to increase sample size. We also plan to analyze the collected video-recorded EHR activities and medical visits to further understand the sources and impact of the extra work.

Acknowledgements: This work is supported by AHRQ funding under #5RO1 HS021290-03n.

References

Predicting Future Anxiety and Depression Diagnoses among College Students Utilizing Electronic Health Data

Jinghe Zhang, MS, Hao Wu, MS, Adrienne Keller, PhD, James Turner, MD, Laura E. Barnes, PhD
University of Virginia, Charlottesville, VA

Introduction
College students responding to the Spring 2014 American College Health Association-National College Health Assessment reported feeling things were hopeless (46%), felt overwhelming anxiety (54%) and more than 80% reported feeling overwhelmed by all they had to do (86%). This critical subpopulation of Americans is facing significant levels of mental health problems, challenging colleges to provide accessible and high quality behavioral health care. However, psychiatric disorders are frequently unrecognized in primary care settings, posing physical, emotional, economic, and social burdens to patients and others. Toward the goal of earlier identification and treatment, we developed and evaluated predictive models that use electronic health record (EHR) data for predicting the diagnoses of anxiety and depression.

Methods
We used anonymous/de-identified archival EHR data collected between Jan 1st, 2011 to Dec 31st, 2014 to develop computational models for predicting anxiety and depression. Our dataset contained 182,845 distinct female patients in the control group (without any mental health diagnoses) and 14,417 patients in the target group (with an anxiety and/or depression diagnosis). Data consist of structured ICD-9 diagnosis codes and demographic data from 9 universities. The ICD-9 codes are aggregated into 283 clusters according to the AHRQ Clinical Classifications Software. We extracted the visit history of patients before their first diagnosis of anxiety/depression in the target group. By selecting patients with at least three visits, there are 88,247 female patients in the control group and 3,553 females in the target group. 80 diagnostic groups are selected using information gain according to their predictive capability of anxiety/depression. Of the selected diagnoses features, “viral infection”, “medical examination/evaluation”, and “immunization and screening for infectious disease” are much more prevalent among patients without mental disorders, while “thyroid disorders”, “headache (including migraine)”, “malaise and fatigue” appears much more frequently among patients diagnosed with anxiety or depression in their later visits.

Instead of traditional predictive models that work well for the average patients, we adopted the idea of personalized prediction according to case based-reasoning where a new problem is solved based on the solutions of other similar past problems. In this research, we computed the pairwise similarity between patients using Jaccard similarity index which captures the occurrences of clinical events in one’s medical history. Three classifiers are built to perform prediction on a patient’s clinical outcome. Model A: K-nearest-neighbors (KNN) with majority voting; Model B: weighted KNN which compares the sums of similarity scores between similar patients and the new patient; Model C: nearest centroid method, in which the centroid is the patient most similar to all other patients in the same class on average and the new patient is assigned the label of the class whose centroid is the closest to this patient.

Results and Conclusions
In our analyses, 3,500 patients are randomly sampled from each group to eliminate the impact of class imbalance. All the classifiers are trained and tested using 10-fold cross validation. The values of k in KNN models are [1, 5, 9, 13, 17, 21]. The performances of those models are evaluated using accuracy, sensitivity, positive predictive value, specificity, and negative predictive value. Model A and B achieve comparable performance with little difference. Overall, the optimal classifier is the weighted KNN when k is 20 and achieving an overall accuracy of 66.3%, a positive predictive value of 69.9%, and a recall of 57.4%.

The developed models could be used to aid clinicians in identifying patients at a high risk of anxiety and depression. However, one limitation is that visits are treated as unordered sets without consideration of potentially important sequential information. Future work will consider sequential modeling approaches. The granularity of the EHR data is another limitation of the current work in terms of its relevancy to anxiety/depression as well as how clinicians use this information to make diagnoses. Practical tools supported by the algorithms in this research could be developed to help students understand their risk of anxiety or depression according to their medical history.

References
Evaluating Efficient Clinician Utilization of Electronic Health Records

Yongkang Zhang, MS¹, Marie Krousel-Wood, MD MSPH¹,², Richard V. Milani, MD²,
Allison B. McCoy, PhD¹,²
¹Tulane University, ²Ochsner Health System, New Orleans, LA

Introduction

Hospitals and other health care providers are putting increased efforts into efficient utilization of electronic health records (EHRs) due to their potential to improve health care quality and safety. Analyzing the actual utilization of EHRs by clinicians therefore is of considerable importance to improve the adoption of EHRs. While prior work has evaluated clinician-reported EHR utilization, studies based on data directly from the EHR about clinicians’ behaviors, which may present stronger evidence, have not been reported.

Methods

We performed this study at a not-for-profit academic health system in southeast Louisiana consisting of 8 hospitals and over 38 health centers in urban and rural settings. Recently, clinicians began using a comprehensive, commercial EHR (Epic 2010, Madison, WI) in both inpatient and ambulatory settings after having used a basic, locally-developed EHR for more than a decade. We analyzed twelve measures of utilization of the new EHR system in four categories, including patient information review, timely completion of activities, order set use, problem list use, and preference list use. These variables were defined as key functionalities by the EHR vendor to be used to assess clinicians’ utilization of EHR. Data were directly extracted from the EHR’s clinical data warehouse. We used descriptive analysis to show the variability of efficient EHR utilization and linear regression analysis to test the clinician characteristics associated with EHR utilization. Analyses were conducted with Stata 13.0.

Result

Data were analyzed for 287 clinicians, consisting of 39,421 encounters during a one month time period, approximately six months after the health system implemented the new EHR. Data show variability in efficient EHR utilization among clinicians. Clinicians highly used the EHR to review the patient information. More than 75% of encounters’ medications, patient history, allergies and problem list were marked as reviewed by clinicians respectively. Clinicians also used the EHR frequently to complete activities in a timely manner. We found that 72.30% of the results-related messages were reviewed in less than 12 hours from being received and 55.41% of patient call messages were reviewed in less than 24 hours. Further, 64.22% of encounters were closed on the same day. Clinicians entered 40.11% of encounters diagnoses onto the problem list, and 69.88% of orders were placed from a preference list. However, the use of order sets was low among physicians. We found that merely 6.65% of orders were placed from an order set, and an order set was used in only 4.66% of closed encounters for face-to-face visits. Low order set utilization is expected, as order sets are designed for inpatient admission orders and not ideal for outpatient visits. We found the utilization of the EHR was associated with several clinician characteristics. Generally, clinicians working in inpatient setting were less likely to use features analyzed in this study compared to those working in outpatient settings. Primary care clinicians were more likely to use these features compared to medical and surgical specialty clinicians. Clinician age was significantly associated with use of order sets in closed encounters for face-to-face visits; we found that younger clinicians were more likely to use order sets compared to older clinicians.

Conclusion

Clinicians showed variability with respect to efficient EHR use. The EHR was highly used for data access and reviewing messages in a timely manner. Some of the frequently used features are related to Meaningful Use incentives. Order sets were seldom used among clinicians, which may be related to local configuration or training. Clinicians’ individual characteristics are related to the efficient EHR use. Further research is warranted and underway to evaluate the implications of efficient EHR utilization on clinician satisfaction and patient outcomes.
Social Media and Autism Support: Health Information Seeking in Facebook by Autism Patients and Caregivers

Yuehua Zhao, MS¹, Priya Nambisan, PhD²
¹School of Information studies, University of Wisconsin -Milwaukee; ²College of Health Sciences, University of Wisconsin - Milwaukee

Background and Significance
The online communities on social media sites provide an efficient platform to autism patients and their caregivers where they can ask for help and advice from other users, make contributions to others, receive assistance from the group members, and share their experiences in the community. In this study, we specifically focus on what kinds of information is being exchanged in an Autism Facebook group and we will classify them into various categories.

Method
We collected data and we are in the process of analyzing the posts from one of the largest Facebook group focused on autism. In total, 17389 group posts and comments ranged from April 2009 to March 2015 were collected by NVivo 10 and NCapture. Content analysis methods is being employed to evaluate the kinds of information users share in this group. This study will address two research questions: 1) what are the characteristics of the autism group on Facebook? 2) What types of information do users share and exchange within the online group?

Results
From April 2009 to March 2015, there were 7709 initial messages posted by 1395 unique group members in autism group, and 9679 messages replied by 1679 users. The average number of comments and likes for each post is 1.26 and 3.03, respectively, which demonstrated the active interaction and strong enthusiasm within the groups. Based on the statistics, users preferred to express their favor of the posts rather than making comments on others’ messages. The potential reason could be the ease of pressing the “Like” button for users when they saw interesting messages posted on the group wall.

All the posts and comments were then imported to SPSS Text Analytics for Surveys software. In total, 15031 records were processed by the natural language processing tools and eventually classified into 9 topics by manual revisions and modifications through data coding analysis: (1) parenting (2513); (2) social support (2245); (3) education (999); (4) research (588); (5) symptom (709); (6) career development (852); (7) therapy (153); (8) relevant diseases (921); (9) daily life (609).

It is not surprising that parenting issues come up frequently, because autism commonly appears in the first 3 years of life and parents usually trying to cope with the stressful situation were seeking help and support from the Facebook group. Given that one of core features of autism patients is difficulty with social interactions and communication, autistic children often face learning disorders. Therefore, education is a popular topic in this group., 999 messages pertained to education, including special education programs, training skills, tutors, etc.

The “relevant diseases” category had 921 posts and comments from group members and carried 10% weight among the record set. The nature of autism and the fact that is one of a spectrum of pervasive developmental disorders probably contributed to this - a lot of posts inquired about the difference between several different but relevant diseases, such as asperger syndrome, pervasive developmental disorder - not otherwise specified (PDD-NOS) and childhood disintegrative disorders.

Conclusions
Facebook group has been used as an efficient platform by autistic patients and their family members. The group was very active and members provided strong support to each other. The topics discussed by group members reflected users’ concerns about autism. Through the coding analysis, this study did a first level classification of topics discussed in this group and our findings show that autism is chronic disease users were more concerned about dealing with it in their day-to day life than the medical aspects of it.
Understanding the Use of Adverse Events Criteria in Radiation Therapy: A Literature Mining Approach

Yi Zhen¹, MS, Yuliang Jiang², MD, Jackie Wu³, PhD, Lulin Yuan³, PhD, Yaorong Ge¹, PhD
¹Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC
²Department of Radiation Oncology, Peking University 3rd Hospital, Beijing, China
³Department of Radiation Oncology, Duke University Medical Center, Durham, NC

Introduction

Accurate scoring and reporting of adverse events in radiation therapy are important for radiation toxicity monitoring and treatment quality improvement. Standardized criteria for scoring adverse events after radiation therapy facilitate normalized evaluation and reporting terms of adverse events and allow accurate analysis of toxicity results at large scale and across multiple studies. Three of the most commonly used criteria RT clinical studies are the Common Terminology Criteria for Adverse Events (CTCAE), the Radiation Therapy Oncology Group (RTOG), and the Late Effect Normal Tissue Task Force (LENT)/subjective, objective, management, analytic (SOMA). These standards have been revised multiple times in recent years. In particular, CTCAE is strongly promoted as the standard for adverse event reporting. However, there is a lack of studies to understand how the various standards have been used in RT clinical studies. This project aims to develop a text mining based method for investigating the usage trend of the three most commonly used adverse events criteria in published literature. In this report, we will address three specific questions: (1) the portion of studies that use each standard; (2) the trend of usage in recent years; (3) the types of adverse events that are reported by each standard.

Methods

We analyzed 531 published articles from January 2010 to December 2012. A total of 668 articles were found in MEDLINE with a search strategy that organized the search terms in following three groups: “radiotherapy, chemoradiotherapy, brachytherapy”, “toxicity, side effect, adverse event, adverse reaction, complication”, and “CTC/CTCAE or LENT-SOMA or RTOG, RTOG/EORTC”. We exclude 104 articles due to inadequate information, and 33 duplicated articles. All selected full articles were downloaded and extracted into text files for analysis. The basic workflow of our project consisted of data preprocessing, statistics calculation, and classifier modeling. We separately used regular expression and Naïve Bayes Classifier to categorize these 531 published articles based on the adverse events criteria applied. In parallel, a radiation oncology physician reviewed all the articles and produced the usage trending data manually. Finally, the manual results were used as a gold standard to assess the text mining results using both approaches.

Results

Based on regular expression analysis, we identified 320 articles using CTCAE, 311 articles using RTOG, 49 using the LENT-SOMA, and 140 using a combination of two or three criteria. From the Naïve Bayes Classifier, we found 340 article using CTCAE, 303 articles using RTOG, 28 using the LENT-SOMA, and 180 using two or more criteria. Compared to the results labeled by the domain expert, the results from the regular expression method reported comparable counts and similar overall usage trends. The results from the Naïve Bayes classifier are also similar to that of the domain expert’s, however the usage trends for the combined-criteria group was over-estimated. As seen in the results, the CTCAE is becoming the most commonly used criteria over time. And RTOG is the most used criteria in 2010, but since 2012 CTCAE has taken over as the most frequently used criteria. We notice that lung cancer studies heavily favor CTCAE while the head and neck cancer studies seem to favor the RTOG standard. We also notice that prostate cancer contributed most radiation therapy articles in the study period while the rectal cancer has least number of studies.

Conclusion

The text mining method based on regular expression is effective in identifying the use of various standard adverse events criteria in RT clinical studies. The analysis shows increased use of CTCAE overall and especially in lung cancer, but also continued use of RTOG in head and neck cancer studies.
Medication Adherence to Oral Hypoglycemic Agents and Hospitalization Cost in Medicaid Patients with Type 2 Diabetes

1Vivienne J. Zhu, MD.,MS. 1Wanzhu Tu, Ph.D, 1Marc B. Roseman, 2J. Marc Overhage MD.,PhD.
1Regenstrief Institute, Inc. Indianapolis, Indiana and 2Cerner, Malvern, PA

Objective: Medication adherence is a major factor impacting patient health outcome and utilizations. Using real world clinical data, we analyzed the correlation between medication adherence to oral hypoglycemic agents (OHA) and the cost of hospitalization related to type 2 diabetes (DM2).

Methods: A sample of Medicaid patients with DM2 who had OHA dispensing records between January 1, 2001 and December 31, 2008 was identified in the Indiana Network for Patient Care (INPC). OHA adherence was measured by the proportion of days covered (PDC) in every 6-month interval starting from the patient’s first OHA dispensing event and further dichotomized using a conventional cut-off point (0.8). The relationship of OHA adherence to healthcare cost during the subsequent six month interval was examined. Healthcare cost included cost of DM2 related hospitalizations and OHA medication expenses under the Medicaid coverage. Mixed effects linear models with repeated measures were used for analyses. Selected patient demographic and clinical factors were included to control the possible confounders.

Results: Among 8,889 eligible patients, 8,288 were OHA non-adherent patients (67.8% female, 76.2% white and mean age of DM2 onset 45.5 years) and 601 were OHA adherent patients (62.6% female, 84.3% white and mean age of DM2 onset 49.4 years). In the non-adherent group, the mean OHA medication expense was $62 and 2,856 (34.5%) patients had at least one hospitalization with an average cost of $6,179. In the adherent group, the mean OHA medication expense was $132 and 200 (33.4%) patients had at least one hospitalization with an average cost of $5,459. The leading comorbidities were hypertension and ischemic heart disease in both groups. After adjusting for confounders, adherence to OHA was associated with lower hospital cost (β = -416, p<0.0001) and higher OHA medication expenses (β =256, p<0.0001). Other factors significantly affecting hospital cost were age of DM2 onset (β =11.1, p<0.0001), female gender (β =-135, p=0.003), ischemic heart disease heart disease (β =701, p<0.0001), stroke (β =536, p=0.006) and hypertension (β =-125, p<0.0001).

Conclusion: In our observational analysis of Medicaid claims data, healthcare costs were significantly lower in patients with DM2 who were adherent to their OHA when compared with those who were non-adherent even after accounting for the higher medication costs.
Facebook and depression: How people with depression use Facebook to manage their depression.
Maryam Zolnoori, Priya Nambisan, Timothy B. Patrick
zolnoori@uwm.edu

Introduction
Internet has provided efficient and effective platform for accessing/sharing health information among patients and caregivers. The recent study conducted by Moorhead, Hazlett, Harrison, Carroll, Irwin, Hoving 1 showed that 61% of adults go online to search for health information and 39% use social media to communicate with patients with the same experiences. To respond to the needs for health information, several applications were created on web 2.0 to allow generating and sharing contents by healthcare consumers. Applications such as virtual content communities (eg, YouTube) and social networking sites (eg, Facebook and twitter) deal with two dimensions of user experiences including cognitive and emotional dimensions 2.

Nambisan 4 discusses four dimensions of social media experience that attract healthcare consumers and facilitate sharing health information. “The four dimensions are pragmatic experience, empathic experience, sociability experience, and usability experience”. Facebook is a social media considering these four aspects, resulting in establishment of several self-management support groups among patients, particularly with chronic diseases.

Research purpose
About 9 percent of American shows symptoms of desperate, regret, and sadness that result in depression (CDC 2014). An initial search in Facebook shows that more than 10 self-management groups (English language) created by Facebook members for depression management. The cumulative number of depression groups members (just in English language) is around 40,000 members. Members usually share their feelings towards their life and their family and friends, and usually look for groups’ member sympathy or support. Studies have found that people with depression are attracted to these groups as they may be getting the kind of understanding and support from group members who went through a similar situations 3. While there are plenty of studies that have looked into online depression groups, there haven’t been any that looked at analyzing posts in Facebook depression groups. The purpose of this study is to analyze posts in Facebook depression groups, to classify them into pragmatic, empathic, and social categories. In addition, nature of comments left for post and their relationships with post are analyzed.

Research Methodology
For this study five groups with around 30000 members are considered. Content, questions, and comments that people posted in these groups have been collected using Facebook API.

Latent Dirichlet Allocation (LDA) tool has been used to model topics that group members posted. LDA has been trained based on algorithm suggested by Hong, Davison 5. Topic modeling is applied to explore purposes and expectations of members from the virtual communities of depression diseases. In the first stage, nature of the posts will be analyzed to understand whether posts are (1) pragmatic (information seeking about symptoms or treatment methods) to manage diseases or (2) empathic (seeking sympathy or attention to cope with emotional gaps); and (3) sociable (extending social network by communicating with supportive people experienced the same medical conditions). In the second stage, the nature of comments left for each posts are analyzed and subsequently divided into three categories: Positive (supportive), negative (unfavorable), or neutral. In addition, relationship between posts and comments are analyzed. Detailed findings will be reported.

Preliminary result
Preliminary result of this research shows that people with depression in Facebook depression groups are more likely to experience empathic or sociability aspects than pragmatic aspect. In fact, group members are more willing to post emotional gaps in their life (such as loneliness), depression symptoms such as self-loathing, or sleep/food disorders explicitly or implicitly and subsequently seeking supportive and sympathetic comments. In addition, they would like to extend their social networking by communicating with group members who experienced the same medical conditions. Moreover, unlike other health communities, members are less inclined to ask questions about treatment solutions, such as medications or brain training, which are common treatment methods in healthcare facilities.

Research Implication
Result of this research will help healthcare organizations to understand coping mechanism of people with depression to manage their disease in virtual environment. In addition, it will enable healthcare researchers to design interventions consistent with needs of patients with depression. Designing proper interventions help healthcare organizations to improve level of patient activation and patient engagements among depressed patients, and consequently, transmit clinical services from determining “what is the matter?” to discovering “what matters to a specific patient with depression disease” 8.

References
Using a Web-based Clinical Pathway and Computerized Order Set to Improve Efficiency of Care for Asthma in a Pediatric Emergency Department

Joseph J. Zorc, MD, MSCE, Ruth Abaya, MD, Eva Delgado, MD, Richard Scarfone, MD.
Department of Pediatrics, The Children’s Hospital of Philadelphia and Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

Background: Evidence from clinical trials suggests that increasing use of metered dose inhalers with spacers (MDIs) to treat mild-moderate asthma increases value of care by reducing length of stay (LOS) and improving resource use. Few studies have evaluated effective ways to use clinical decision support (CDS) in an electronic health record (EHR) to support this recommendation in the Emergency Department (ED).

Aim Statement: For mild-moderate acuity ED asthma patients, to use a web-based pathway and EHR order set to increase use of MDIs and reduce LOS.

Methods: A baseline review focused on discharged patients ≥ 2 years with an asthma diagnosis and a non-emergent triage assessment (Emergency Severity Index (ESI) 3 or 4). This review identified a median ED LOS of 3 hours and that 40% received an hour of continuous albuterol (CA). 90% of patients received an MDI at discharge, representing duplicated resources. Our multi-disciplinary team developed an intervention to increase use of MDI instead of CA for acute treatment using the following interventions: 1. Redesign of the ED Asthma pathway (www.chop.edu/clinical-pathway/asthma-clinical-pathway) and order set to provide CDS recommending MDIs for ESI 3/4 patients and encourage early discharge planning. Key elements of the order set included: active suggestion to providers at time of ordering, and discrete, actionable options based on ESI level. 2. Adding conditional order for the Respiratory Therapist to reassess and repeat MDI q20 minutes x3 until patient reached mild assessment. Process and outcome measures were tracked; balancing measures included admission and revisit rates for target patients and higher severity asthma patients.

Outcomes: The interventions were implemented and assessed in 2,667 ESI 3/4 asthma patients treated from July-December 2014. For target ESI 3/4 patients, CA use decreased immediately from 40% to 13%, and the percentage discharged in < 3 hours increased from 53% to 63% (both p<0.05 using process control measures, see Figure). Measures for more severe patients did not change, although overall ED LOS for all asthma patients decreased by 22 minutes after the intervention (p<0.05). Monitoring continues to show sustained change, and ongoing interventions are occurring.

Conclusion: Preliminary data show that a revision to a web-based ED asthma pathway and EHR order set using CDS principles reduced CA treatment, and improved efficiency for the care of mild-moderate asthma patients.

KaryoViz: Designing A Karyotype Visualization Platform for Clinical Decision Support

Zachary B. Abrams, BS\(^{1}\), Satyajeet Raje, MS\(^{1}\)

\(^{1}\)Ohio State University, Columbus, OH

Introduction

Data visualization is a critical step in helping researchers better understand the structure and meaning of their data. In the domain of medical genomics cytogenetic data is frequently used to provide visualization at the chromosomal level. Karyotypes are syntactic descriptions of chromosomes, derived from a microscopic visual examination of chromosomes. Karyotype analyses help detect chromosomal structural and numerical defects that can serve as genetic indicators of disorders or diseases. These tests are more readily available than the more specialized sequencing technologies and are thus a more accessible means of assessing a patient's genetics for diagnosis and/or therapeutic planning purposes.

It has long been observed that karyotype analysis can be used as a diagnostic tool in medicine since many diseases have distinguishable patterns of chromosomal aberrations\(^{1}\). The International System for Human Cytogenetic Nomenclature (ISCN)\(^{2}\) is a domain-specific language that records these chromosomal defects through visual microscopic inspection of chromosomes. However, the karyotype information in this format is difficult to analyze using existing computational methods by virtue of their syntactic variability, information density and potential for human error. Thus, the karyotype data remains severely underutilized.

Prior attempts have parsed ISCN karyotypes but have failed to accurately map them to a structured language biological model\(^{3}\).

In our previous work, we propose, 1: a Loss-Gain-Fusion (LGF) biological model (AMIA citation) that provides a standardized machine-readable representation of the karyotype information and 2: a computational cytogenetic platform that transforms ISCN-encoded karyotypes through a process of parsing and mapping into their standardized representations using the biological model. Thus, the platform provides an automated pipeline that can extracts biologically important information from text-based karyotype data in a form that is ready for reuse. The karyotype data can then be utilized for several clinical as well as applications such as visualization, analysis, etc.

In this paper, we highlight the data processing and visual component of this pipeline. We propose a visualizaion tool, called KaryoViz, that will primarily allow clinicians to input the syntactic representation of a karyotype and analyze it in an intuitive interface. The users would be able to navigate through different granularity of the information extracted from the karyotype data. It would also allow the users to visually browse through the Mitelman database, which is the the largest public repository for karyotype data\(^{4}\). Thus, the tool would be instrumental in augmenting clinical decision support by reusing the karyotype data, which is otherwise be inaccessible to clinicians and researchers alike.

Parsing

The parsing process converts natural language into computer language. To transform ISCN-encoded karyotypes (natural language) into machine-readable constructs (computer language), several parsers were generated using ANTLR (ANother Tool for Language Recognition)\(^{5}\) Context-Free Grammar (CFG), encoded using the Extended Backus-Naur Form, was used to write ISCN-specific production rules. This approach was based on the observation that any ISCN karyotype was describable in its entirety using a recursive walk of its morphological composition using a minimal number of CFGs. Several such production rules were created, starting with a fundamental rule for a karyotype and incrementally adding further rules for recognizing the collection of possible patterns for each aberration in that karyotype. Thus, each complex karyotype was morphologically deconstructed into constituent phrases. Finally, computationally readable “parse trees” were constructed using these phrases for downstream usage in the overarching computational pipeline.
Mapping

Mapping the parsed karyotypes into our machine-readable model required the creation of a domain-specific language used in the mapper function. This nomenclature allowed chromosomal aberrations to be rewritten and then redefined into the LGF model. First, the ISCN abbreviation for the aberration was written, followed by the number of elements that would be extracted from that aberration. Each extracted element consisted of the chromosome, the arm of that chromosome, and the band of that chromosome. A colon was used to distinguish biological events. Then, starting with element zero, each element was assigned to its respective part of the biological model. The biological elements were separated by a "," and different abbreviations represented how that element should be classified based on whether that element was lost, gained and/or involved in a fusion event. Thus, ISCN karyotypes were rewritten to purely represent the biological function at each location. Although antecedent events, such as translocation and deletion, were not preserved in the model, the biological effect of those events was represented, allowing for comparison across aberrations with similar biological effects. Each aberration had its own unique classifier in this domain-specific language. Following the use of the domain-specific language to map the karyotypes into the LGF model, we were generated a matrix in which each row represented a karyotype and each column represented a potential cytogenetic band. Let us now look at how the above workflow can be utilized in the proposed visualization tool.

The Loss-Gain-Fusion (LGF) Biological Model

The design of the LGF model was guided by the three major events known to cause deleterious oncogenic effects at the chromosomal level, viz. loss, gain and fusion. The first event is the loss of a gene, which can disrupt the cell cycle or other critical pathways. Deletion is an example of a loss event on the chromosomal level. Deletion of an entire section of chromosomal DNA represents a loss of DNA, thus the genes in that region are not present on that chromosome. The second event is the gain of an additional copy of the gene. This causes additional gene products to be created that could disrupt the balance of pathway regulation in a comparable manner to the loss event. As with a loss event, a gain event alters one component of a pathway and thus can disrupt that pathway's function. Duplication represents an example of a gain event, wherein a segment of chromosomal DNA is represented twice. If the fully intact gene is present within the duplicated region, that gene may express more product than a non-mutated cell. The third event is a fusion event that can produce a chimeric oncogene that disrupts the cell cycle in some capacity. A fusion event requires a breakage event so there could also be a loss of function at fusion points as well as the possibility of a chimeric oncogene. The chimeric gene is formed from multiple genes; this can occur due to a fusion event. It is possible for chimeric genes to alter cell cycle regulation or other pathways that, when disrupted, can lead to a loss of regulation that leads to oncogenic development.

The proposed Loss-Gain-Fusion (LGF) biological model represents the aberrations in the karyotype based on the biological function (loss, gain, or fusion) occurring at that location in the chromosome as opposed to focusing on the event that caused that particular biological function (e.g., deletion, translocation). This places sole computational emphasis on the underlying biological result of the event rather than on the event itself. This allows researchers to compare across events in the analysis of similar functional outcomes in patients without having to rely on similar events causing those functions.

We began by utilizing the nomenclature found in the ISCN language for writing karyotypes. We then created a set of parser rules that allowed us to break down the text form of the karyotype into a tree structure. By traversing the tree and pulling out the tokenized elements representing the cytogenetic events (e.g., translocation and the location - e.g., 17 P 13), we could reassemble the event and location into pairs. Then we developed a mapping language that would assign each possible cytogenetic event to one of three overall biological events: loss, gain or fusion. This allowed us the ability to map each location and the potential event that could occur at that location. This in turn allowed us to represent the text base karyotype as a binary vector where each cytogenetic region has three representations, i.e., the loss, gain, or fusion form of each location. There are 910 cytogenetic locations within this model, thus the binary vector for each karyotype was represented using 2,730 units. This binary vector is standardized, of a fixed size and can be read computationally for several different applications.
Karyotype Transformation Workflow

This workflow (Left) illustrates the process of converting raw text based karyotypes into the LGF binary vector model. The first step is to parse the raw text karyotypes using a set of production rules. This produces a tokenized form of the karyotype, i.e., all text in the karyotype is broken down into individual elements and then examined based on context in relation to other elements. This allows the chromosomal aberration and the affected band to be isolated and classified into the LGF model. The LGF classification is based on the mapping language that classifies each band based on its corresponding aberration. This outputs a vector representing each chromosomal region, where each region has a three bit code denoting what event has occurred at that location.

Input Data

Traditionally, karyotype data are acquired from a patient’s tumor by selecting tumor cells in mitosis. The researcher then maps the chromosomes into chromosomal pairs and arranges them in order. The patient's aberrations are noted in the ISCN format and an ISCN karyotype is written based on the number of chromosomes and aberrations in the patient’s karyotype. An example ISCN karyotype is presented here: 47,XY,del(20)(q13),+21.

Design Requirements

Potential Use Case: User uploads a particular karyotype to visualize

1. To begin, a researcher will upload a karyotype and our system will parse and map that karyotype into the LGF model as previously presented.
2. Then a personalized ideogram is generated and presented to the researcher. This ideogram shows all chromosomes and color-codes those bands involved in losses, gains, and/or fusions. This allows the researcher to quickly identify those chromosomes and/or individual bands of interest.
3. If a researcher wishes to observe a deeper level of granularity on a particular chromosome, they simply select that chromosome and the system zooms in to a more detailed map of that particular chromosome. Each band and sub-band within the chromosome is still color-coded using the same key so the researcher can identify those areas of interest.
4. By simply clicking on a sub-band, a list of all genes present on that sub-band will appear next to the band. This provides the researcher better visualization of their data and thus enhances the understanding of the various gene level changes caused by the cytogenetic aberrations.
This form of data visualization stands in contrast to the current standard to visually examining the karyotype based on the ISCN text string. This system of visualization thus plays to the human strength of visual pattern recognition to determine potentially important elements within the karyotype.

**Design Elements**

*Ideograms:* Karyotype ideograms, commonly used in cytogenetics\(^2\) proved to be far more visually informative than binary vectors. An ideogram is a map that illustrates all the chromosomes and bands within chromosomes. We took this design concept and created an interactive ideogram that was color-coded based on the researcher’s inputted data. This allowed us to visually display a new form of cytogenetic data utilizing a familiar visual system; the ideogram. This helps make the system more intuitive to cytogeneticist who are used to seeing information displayed in this form.

**Implementation Details**

The KaryoViz tool is under active development both as a web application as well as a mobile application. We are using the D3 visualization library for implementation. The parsed Mitelman data is stored in a no-SQL database. The current focus is on finalizing the user interface design. The details of the current demo interface can be found in the supplemental material (\(s1\)).

**Evaluation**

An important part of the development process is testing the UI. The user interface design will be evaluated in two ways detailed below. Both these involve a “cold” evaluation of the interface, meaning that the UI is not necessarily connected to a functioning and fully working system.

*Heuristic evaluation*

The heuristic evaluation will be performed in two parts. First, the interface is to be evaluated by UI experts based on the HCI principles\(^8\). Next, we would design a survey based on these principles for the users of the interface. The platform will be evaluated as a combination of the expert comments and the user responses.

*Cognitive evaluation*

This is a task-based evaluation. An optimal workflow to perform certain tasks is prerecorded. The tasks are designed around the different functions that the user is expected to perform using the interface. A set of test users will be asked to perform the same tasks. The think-aloud protocol will be used as users are performing the tasks. Both the user actions and the responses will be recorded. An evaluator would also note various aspects of the process, such as whether the user could actually finish the task, the deviation from the optimal flow, time required to complete etc. All of this data will then be analyzed to evaluate the user interface. Additionally the eyeball tracking method can also be used as an additional cognitive response measure of the system.

**Conclusion**

The ability to visually inspect genomic information represents a significant step in cytogenetic research. Our proposed Cytogenetic Platform contains a visualization system that spans the scale of observation from the large-scale chromosome level to the individual gene level. It empowers the researcher or clinician to visually target that level of data that best meets their needs. This represents an important visual tool for cytogenetic researchers.

**References**


The EH Tracker: Using Dynamic Environmental Health Data for Improved Personal Health Decision-Making

Juandalyn L. Burke¹, MPH, Dae Hyun Lee¹ and Ryan C. James¹
¹Biomedical Informatics and Medical Education, University of Washington, Seattle, WA

Abstract

Although climate change and environmental health information have been widely dispersed to the public through various channels of communication, from environmental advocate groups to mass media productions, little is known by the public about how the impact of dynamic environmental health hazard data affects their individual health. Existing environmental health applications consist of predominantly pollen allergy and weather forecast based systems, only provide the public with a limited set of environmental health information, lack recommendations for ways to reduce the public’s exposure to predominant pollen allergens, and do not include any data on air pollution levels. In our solution, we aim to provide individuals with user-specific environmental health hazard data using a web-based and mobile application called the Environmental Health (EH) Tracker. The EH Tracker design incorporates actionable recommendations that reduce exposure to pollutants and pollen allergens, and offers features such as electronic health record integration and healthcare provider communication channels. Our design includes machine learning algorithms that provide users with environmental hazard feedback summaries, recommendations for reducing exposure, and healthcare provider notifications triggered by user experiences. Thus, we believe that the EH tracker will aid in reducing individuals’ exposure to environmental health hazards and reduce their chances of having an allergic reaction and/or worsened existing health condition such as chronic and non-communicable diseases.

Introduction

With rapidly evolving weather patterns and increasing levels of air pollutant emissions, public safety has become an enormous challenge. How does the public stay informed about these challenges? How do they discover the types of environmental hazards are affecting their health? How do they protect themselves from being exposed to these hazards? At the 2014 Climate Change Summit President Obama stated that the “climate is changing faster than our efforts to address it...”, and addressed the need to have more technological innovation in an effort to reduce exposure to environmental health hazards¹². The link between disease susceptibility and environmental hazards, such as air pollution which causes climate change, has been well-researched over the past sixty years³⁵. Most notably, in 1948, the national coverage of the citizens of Donora, Pennsylvania being diagnosed with respiratory and cardiovascular illnesses as a result of a “lethal haze” entrapping air pollutants in their community brought about government mandates such as the Clean Air Act of 1970⁶. However, air pollution remained a persistent issue due to the consistent release of air pollutants by power plants and motor vehicles¹. According to the Center for Disease Control, respiratory illnesses such as asthma have become a “national epidemic”, affecting one in twelve people in the United States with a growth rate of 28% from 2001 to 2011¹³. Among various pollutants affecting our air, land, and water, air pollutants are more likely to vary periodically, unexpectedly exposing individuals to high levels of toxic pollutants. Now more than ever, it’s imperative that we build innovative solutions that predict the effects of environmental health hazards and prevent the public from experiencing their negative effects.

Currently, there are several environmental health apps that inform users about city-wide data related to pollen count and weather forecasting¹¹¹³. Each of these apps allow their users to log symptoms in a daily diary. However, the diary has limited functionality for exporting and sharing data with other users or medical professionals, such as their doctor or nurse⁶. Moreover, the existing environmental health apps inform their users of the environmental health hazards present in their community by displaying range values to indicate severity based on the average person¹³. However, these apps do not give users a personalized medical perspective detailing how environmental health hazards affect their individual health. Such an app, based on personalized medical data, can deliver actionable information to users, provoke specific, positive behavior changes on an individual basis, and reduce exposure to harmful effects at the population level.

Thus, we plan to design and develop the Environmental Health (EH) Tracker to provide users with: (1) a more accurate, user-specific environmental exposure and health-risk assessment based on machine learning algorithms, and (2) channels of communication with healthcare providers facilitated by user-specific environmental health
Alternative Solution

As an alternative solution, we considered designing a Geographical Information System (GIS) to present environmental information to susceptible users. Supplementary Figure 8 shows the overall concept of the alternative solution. The preprocessing step collects air pollution data from various sources and feeds the data into a machine learning algorithm that trains an environmental health risk prediction model. The following data were identified as requirements for training models: pollen count, air pollution level, and traffic congestion patterns.

When using a machine learning approach to derive accurate prediction models, it is crucial to prepare with the right training data. Past events can be a great source of data because the outcome is known. Variables in these data sets not only represent pollution levels such as pollen count and dust level, data for confounding variables such as wind speed and precipitation can be used for training as well. In addition, patient outcomes can be obtained from public health surveys such as NIHS (National Health Interview Survey) or BRFSS (Behavioral Risk Factor Surveillance System). We planned to integrate these heterogeneous datasets into a single dataset for training purposes.

With Principal Component Analysis (PCA), distinctive features in the training dataset can be extracted and used for training multiple health risk prediction models. Selecting the important features to analyze within the training dataset is a crucial step in training such models; it not only effects training efficiency, but also each model’s level of prediction accuracy by avoiding over-fitting. We contemplated two ways to use the models after they were generated. First, the model with the highest accuracy would be chosen. For example, the model that performed the best when predicting pollen outcomes would be used to predict pollen. Second, each model would be integrated into an ensemble model, or a meta-predictor would be used to select the best model based on the pollutant under analysis.

However, we identified shortcomings for the alternative solution through persona and scenario-based evaluation sessions. First of all, the design could only manage one profile at a time. Second, although it could convey multiple environmental hazards for a single location, the system could not present any actionable information to users. Lastly, it could not be customized for individual users. Through a deeper understanding of the system’s stakeholders we identified that user-specific information is a crucial feature for the system because the severity of the symptoms caused by poor air quality highly depend on an individual’s pre-existing conditions. In addition, allergic reactions tend to have high specificity to allergens. Therefore, we concluded that this “one-size-fits-all” scoring system was inadequate for an intervention that provides users with meaningful, user-specific health risk prediction and prevention information. After analyzing the alternative solution’s strengths and weaknesses we arrived at our proposed solution—the EH Tracker.

The Solution

Our proposed solution has three main objectives: 1) to present personalized environmental health information to users in an intuitive, user-friendly interface, 2) to provide a tool for individuals to avoid environmental factors that adversely affect their health, and 3) to provide a useful environmental health supplement for clinical visits. We identified these objectives when we conducted background research, which included the identification of stakeholders, the development of personas and scenarios, and a participatory design session conducted with a group of our peers. The objectives also represent a set of gaps that currently exist in environmental health apps on the market today.
Figure 1. The EH Tracker system architecture. Environmental and health data (left) is combined with user-specific data (top) and fed into EH Tracker features (bottom) to provide users with actionable environmental health information to improve their day-to-day lives. These experiences can then be sent to other platforms (right).

Key features

In order for the EH Tracker to fulfill the above objectives, users will need to engage with it and the underlying data in meaningful ways. We aimed to abstract away the ‘big data’ and computational complexity that drives the EH Tracker interface and present the user with a limited set of easy to use features that improve their quality of life. These features, as seen in Figure 1, include: an alert system, an end-of-day user feedback mechanism, a health history view, a profile page, and a home page summary view.

Alerts (Supplementary Figure 1a). When changes in the environment will have an adverse effect on an individual’s health, the individual needs to be informed with actionable information enabling them to avoid harm. If users are required to check for environmentally important information themselves they may check late or not at all, both of which could cause the user serious environmental harm. To avoid these issues we will present users with important information as soon as they need it. Unfortunately, alerting in healthcare has acquired a bad stigma. Alert fatigue associated with Electronic Health Records and other systems has become a pervasive issue. To address potential alert fatigue we designed the EH Tracker alert system to alert based on user preferences. Users will set alert settings when they create their profile and can change alert settings at any time in the future. To determine the appropriate set of alert options we plan to perform an extensive user study. For more information about the user study please see the future work section.

User feedback (Supplementary Figure 2). The EH Tracker’s health risk prediction service is built on top of a machine learning algorithm that is fed user feedback as input. The prediction service is the backbone of the EH Tracker application. It helps users avoid health issues and discover their inherent susceptibility by providing users and their health professional with data they can use to identify and treat symptoms. Because user feedback is such an essential part of the EH Tracker service, we designed a user friendly feedback mechanism that captures information about a user’s health when it is needed. For example, if a user is in an environment with a high pollen count the feedback mechanism may ask them about how they are feeling and which medications, if any, they have or will take. The machine learning algorithm then uses this information to predict the user’s susceptibility to pollen. If they are susceptible the app will notify the user of any future pollen health risks, and summarize their health history information for healthcare professionals.

User Profiles (Supplementary Figure 3). Creating user profiles is essential to the customization of the EH Tracker user interface. Profiles enable the user to have control over much of the information that is displayed after they log in. The creation of the profile page specifically identifies the user’s name, contact information, alert settings, and current medications. It also confirms the sharing of information with other users, and the option to view their own current health status. In addition, the profile page is specific to the type of user using the system. We have identified two main users: direct and indirect. The direct user is in direct contact with the system, such as a student, caregiver, or businessman/businesswoman. These direct users serve as application owners, and are the primary data
susceptible to environmental pollutants. These users have authorization to customize the features displayed in the EH Tracker user interface and are primary contributors to the daily feedback data. In contrast, the indirect users are consumers of information. These users include children or other persons with a caregiver who are unable to use the EH Tracker system. Another user who serves as both a direct and indirect user is the physician or healthcare provider. The physician or healthcare provider profile is able to integrate users’ electronic health record (EHR) information with the EH Tracker’s environmental and user feedback data to provide a comprehensive diagnosis and healthcare recommendation for users. In addition, the healthcare provider’s profile page allows for better patient-provider communication by supporting direct communication with their patients via email, phone, or in an open chat.

**Home page** (Supplementary Figure 1b). The EH Tracker homepage is the first page the user views when opening the app. The homepage serves as the user's brief look into the current status of environmental health hazard levels, access to other page views such as their current health status, links to the prevalent pre-dominant pollen allergens information, notifications for updated information from their healthcare provider or loved one (see Supplementary Figure 1b). Environmental hazard information related to air pollution such as ozone and particulate matter (PM2.5) air quality index ranges are indicated by colored clouds reflecting the Environmental Protection Agency’s Air Quality Index ranges of 0, “Good” with a green bar to 500, “Hazardous” indicated by a maroon color. The predominant pollen allergy type is also specified using a photo of the pollen plant species, and a gradient bar indicating level of severity on the impact of an individual’s health. Major features of the homepage also include the ability for physicians to assist with prescribing and notifying patients about recommended medications, and the ability for users to briefly view actionable information they may be able to include in their day, such as running recycled air in their car or wearing a mask when they go outside during certain times of day.

**Implementation and Evaluation plan**

![Timeline for implementation and evaluation](image)

**Figure 2.** Timeline for implementation and evaluation. Each development phase has implementing features, conducting usability test, releasing, conducting post-user survey, and fixing identified defects. See Supplementary Figure 4 for the detail.

To continually help individuals that are susceptible to environmental pollutants we developed a detailed implementation and evaluation plan for our solution. To date, much of the work has been completed. At the beginning of April we performed in depth background research to converge upon a problem statement and set of primary, secondary, and tertiary stakeholders. To gain a better understanding of these stakeholders, we leveraged design literature to guide the development of three personas (see Supplementary Figure 5-7) and several scenarios (see Supplementary Figure Box 1). These artifacts were then used to aid in the development of a detailed evaluation plan.

Our evaluation plan (see Supplementary Figure Box 3) describes design characteristics, methods, and procedures for evaluating our first design iteration—the product of our first development phase. We plan to revisit the plan during the first development phase to address any changes between now and then. In this day and age users’ needs can change quickly, highlighting the importance of iterative design. We plan to use this methodology to continually monitor and understand user needs, and maintain confidence that we are building the best solution for individuals susceptible to environmental pollutants. In addition, we plan to conduct a similar evaluation as part of each
following development phase. Continual evaluation will help our research identify design problems, increase understanding of our users’ needs, and increase the value added to the collection of environmental health research artifacts used by other researchers in this area.

**Conclusion**

In this abstract we present the EH tracker—a cross platform application that provides user-specific health information to individuals at risk of environmental pollutants by leveraging user interactions, machine learning algorithms, and multiple dynamic environmental health data streams. EH Tracker’s ability to present users with actionable data that’s specific to their environmental health needs is what sets it apart from existing apps, and what makes it essential for individuals affected by environmental pollutants. We believe that EH tracker is better positioned to help consumers make environmentally conscious health decisions. By merging user-specific data with environmental data, the system will present more accurate predictions accompanied by actionable information based on user-specific trends. In addition, EH tracker will aid in effective patient-clinician communication by exporting user feedback and health history in a structured format. With detailed information about environmental hazards and patient symptoms, clinicians can use EH Tracker summaries to facilitate understanding symptoms, formulate diagnoses, discuss personalized treatment options, and improve patient care.

**References**

Supplementary Figure 1. a) Alert notification b) The EH Tracker Homepage.
**Supplementary Figure 2.** User feedback page. If the prediction model identifies the possible relationship between environmental risk factor and the disease, it will trigger alert with actionable information (bottom).

**Supplementary Figure 3.** The information view of loved one. It represents the current symptoms and the person who checked with the timestamp (top), medication and photo of symptoms (middle) and the list of people who have access to the profile (bottom).
Supplementary Figure 4. Detailed implementation and evaluation plan
Supplementary Figure 5. Persona of working mom

Demographics
- **Age:** 33 years old
- **Marital Status:** Married
- **Children:** 1 kid (8 months old)
- **Pets:** none
- **Location:** Chicago, IL -> Atlanta, GA
- **Occupation:** Public Health Researcher
- **Income:** $65,000
- **Education:** Spelman College and Univ. of Michigan

Health Status
- **Allergies:** Pollen, Dust, and Air Pollutants
- **Diseases:** Asthma
- **Medication(s):**
  - Advair or Symbicort; Albuterol; Systemic corticosteroids
  - (depending on severity of symptoms)
- **Physical Activities:** not physically active; trying to avoid any asthma attacks

Her Life
- Frequently checks email
- Daughter has frequent ear infections while at daycare.
- Loves to dance and garden, but doesn't have the time or energy
- Passionate about career but requires her to work long hours, and occasionally travel.

Environmentally Conscious
- Knows that pollution affects her asthma.
- Does not know how to better her health other than through taking medication.
- Checks weather updates to prepare proper attire for herself and her child
- Does not have time to become well-versed in environmental health; relies on clinical information.
Supplementary Figure 6. Persona of young active user
Supplementary Figure 7. Persona of frequent traveler
Supplementary Figure 8. Overall design for the alternative solution
Supplementary Box 1. Value-sensitive scenarios for the EH Tracker

Scenario 1: The Parent
Claire is a helicopter parent who does not have time to check on her 2 year old child while at daycare. Her 2 year old has increasingly severe reactions such as runny nose, puffy, red eyes, and the oozing of mucus coming out of her ears. These symptoms are also accompanied by fevers and frequent ear infections. When she inquires with the daycare provider about when and how these symptoms have occurred, the daycare provider tells her she is unsure. The daycare provider then explains that if her 2 year old continues to exhibit these symptoms, she will have to take her child to the doctor and her child will not be able to come back to the daycare without a cleared doctor's notice. Claire usually is unable to take time off of work due to budget cuts that left the office short staffed. Thus, she feels hopeless in finding ways to track and prevent her child's allergic reactions. She does not want to jeopardize her job by having to take off of work. Claire decides to call her healthcare provider to see if there is another option she can use to help monitor her child's allergic reactions. The doctor tells her to download the environmental health tracker app which may better assist her in managing her child's allergic reactions. The mother shares this information with her daycare provider to ensure that they are well-informed about the possibility of her child exhibiting symptoms today and the appropriate medications to use. Once the child is at daycare, the mother is able to be alerted by the app when the daycare provider selects different symptoms her child may be presenting and/or if any medication was taken. This information is shared with the pediatrician, as well.

The next day, as she is cooking breakfast, she is alerted by a message on her phone. The EH Tracker suggests that the ozone and particulate matter levels, pollen count and type of pollen may affect her and her child that day. From the air pollution levels, the app makes the suggestion to make sure she uses re-cycled air when running her car's air conditioning. From the pollen forecast, the app then makes suggestions about possible medications like Benadryl that the child should take if an allergic reaction occurs. Given the side effects on the medication, it also tells her when may be the appropriate time to take the medication (such as 30 minutes before nap time). The mother shares this information with her daycare provider to ensure that they are well-informed about the possibility of her child exhibiting symptoms today and the appropriate medications to use. Once the child is at daycare, the mother is able to be alerted by the app when the daycare provider selects different symptoms her child may be presenting and/or if any medication was taken. This information is shared with the pediatrician, as well.

Scenario 2: The Traveler
Kenneth receives an email on his phone with the time and location of his upcoming meeting. In an effort to stay organized and have a reminder to attend the meeting, he clicks on a hyperlink in the email which adds an event to his calendar. After the event is added, the Environmental Health (EH) Tracker pop-up window appears on his mobile device and asks him if he wants to add any EH Tracker alerts to his calendar for that day. Simultaneously, the tracker also provides more information about the predicted discomfort he will suffer based on prior data about his health conditions, including previous interactions with environmental pollutants. Since the EH tracker predicted a high discomfort score, he decided to add the EH Tracker alert to his calendar. The night before the meeting the EH Tracker alert popped up with a more accurate, high discomfort score. Unfortunately, he ran out of medication, so he opens the app and clicks on the prepared button. The EH Tracker launches the navigation service and provides locations where he can buy his medication near his travel route. He decides to buy Zyrtec at the Walgreen next to his meeting location. On his way back home after work, the EH Tracker pops up on his phone to ask him about his experience that day. He provides specific information about his experience that day, not by typing, but by selecting choices. He indicates that (1) he had a swollen nose, (2) it was not significant enough to take his medication, and (3) that he may be allergic to a specific type of pollen that he suspects is causing his symptoms. The tracker also provides a way to log the estimated exposure time for the pollen.
Scenario 3: The Student

Dave wakes up in the morning and heads to the bathroom to wash up and brush his teeth. While he brushes his teeth he thinks about his schedule for the day--school, then home. As he finishes up in the bathroom he spots the allergy medicine he's been trying out next to the sink, but ignores it since his day will be spent primarily indoors. He heads back to his room, gets dressed, grabs his things and prepares to head to school. When he picks up his phone he notices an alert from the environmental health tracker app--it's notifying him of a high pollen count in a nearby city. He wonders why he received the alert, so he selects it to find out more. To his surprise the app remembered he had a soccer game in a nearby city, and suggested he take his medicine as a precaution since last time he was in an area with high pollen count he suffered from a runny nose and frequent sneezing. Feeling well informed, he grabs his medicine before heading to school.

Later that night Dave receives another alert asking him how his day went. He selects the alert, and when the app opens he presses the smiley face, indicating he had an allergy free day. The app then suggests with high confidence that Dave is allergic pollen, and shows him his past responses that led to this conclusion--on several occasions Dave pressed the sad face when he was in areas with high pollen counts. Dave appreciates that the app remembered his responses and used them to help him stay healthy over time. He is confident that the app's notifications and apparent machine learning algorithm will help him improve his health and quality of life. From now on he plans to check the app first thing when he wakes up.
**Supplementary Table 1. Direct stakeholders for EH Tracker**

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Track their symptoms, tag their geographic location, and access the most up to date pollen and air pollution data. May use system to upload pictures and video of their reactions.</td>
</tr>
<tr>
<td>Teacher</td>
<td>Create/import profiles for students with allergic reactions that contain important information, such as allergens and medication. Check pollen and air pollution data to prepare for and prevent allergic reactions of children in the classroom</td>
</tr>
<tr>
<td>Parent</td>
<td>Track child's geographic location over time. Monitor child's allergic reactions while away (e.g. at work). Communicate with their child's teacher.</td>
</tr>
<tr>
<td>Traveling Employee</td>
<td>Track environmental factors in the areas that they work. Use knowledge of environmental factors to prepare for meetings and travel.</td>
</tr>
<tr>
<td>Athlete</td>
<td>Prepare for athletic events (e.g. soccer game, track meet) despite their location. Set reminders and alerts to stay informed.</td>
</tr>
</tbody>
</table>
| Person affected by allergies or other chronic conditions | People affected by allergies will be able to determine if they need to…  
  - Take certain medications (i.e. in order to ensure the ability to tolerate certain allergens)  
  - Wear a mask  
  - Stay indoors during certain times of day  
  - etc. [http://www.oprah.com/health/How-to-Protect-Yourself-from-Air-Pollution](http://www.oprah.com/health/How-to-Protect-Yourself-from-Air-Pollution)                                                                                                                                                    |

**Supplementary Table 2. Non-target use stakeholders for EH tracker**

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharmaceutical companies</td>
<td>Analyze the number of people experiencing discomforts due to the environmental factors, and identify new risk factors that a drug needs to deal with.</td>
</tr>
<tr>
<td>Drug Store</td>
<td>Predict the market demand of their drugs and competing companies'</td>
</tr>
<tr>
<td>Insurance Companies</td>
<td>Adjust insurance rates based on an individual's susceptibility</td>
</tr>
<tr>
<td>Scientist or Researcher</td>
<td>Potentially use the environmental health tracker data to make predictions or simulate potential areas that will be most affected, sub-populations/groups of people to be affected, and/or conduct PSAs to make people aware of how the environment is affecting their health.</td>
</tr>
<tr>
<td>Thieves</td>
<td>Potentially know when users are not at home. Phishing/Masquerading attack (impersonate user to gain access to their data)</td>
</tr>
</tbody>
</table>
**Supplementary Table 3. Indirect stakeholders for EH tracker**

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politicians</td>
<td>Public policy advocates could create laws to change thresholds of emissions for air pollution.</td>
</tr>
<tr>
<td>Government Agencies such as the Federal Drug Administration (FDA)</td>
<td>May use this information to make more enhanced medications to support those who have frequent or severe reactions to pollen or air pollution.</td>
</tr>
<tr>
<td>Environmental Advocates</td>
<td>May use data to support their agendas (e.g. env clean up)</td>
</tr>
<tr>
<td>Real Estate Companies</td>
<td>Could use data to sell properties in areas with low pollution</td>
</tr>
<tr>
<td>News Reporters and Weather Reporters</td>
<td>Use data to report news</td>
</tr>
<tr>
<td>Companies/Areas producing pollution</td>
<td>There may be greater pushback by the users of this system and other influenced entities to clean up and produce less pollution</td>
</tr>
<tr>
<td>Schools and Daycares</td>
<td>May not want to have another task of recording more information about children on a daily basis (need to be careful to make sure this doesn’t adversely affect workflow).</td>
</tr>
<tr>
<td>Doctor</td>
<td>May use the data to help diagnose and track their patient's health over time, and use the information to understand the underlying cause of their reactions.</td>
</tr>
<tr>
<td>Child</td>
<td>May remind parent to use the system. May learn about environmental health issues and be more knowledgeable about ways to reduce air pollution, and live in a cleaner environment</td>
</tr>
<tr>
<td>Possible future trends using EH Tracker</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td></td>
</tr>
<tr>
<td>• Public service announcements (PSAs) may be posted in the media across electronic billboards/signs, or in television commercials that uses information from the EN Tracker. This would especially be important for those who do not have access to the technology the option to improve their health.</td>
<td></td>
</tr>
<tr>
<td>• Companies may change work hours to times when in order to minimize their employees exposure to pollutants.</td>
<td></td>
</tr>
<tr>
<td>• People will expect to easily find the closest drug supply stores close to their travel route so that they can be prepared for the foreseeable discomfort. Also, the travel route calculation takes the environmental factors into account in a way that people are less exposed to the disease-causing factors.</td>
<td></td>
</tr>
<tr>
<td>• There may be an expectation to have the EH Tracker incorporate seasonal weather patterns, users' health data, and pharmacy information to improve the accuracy over time. Moreover, there may be a need to customize a mathematical model that ensures the privacy of certain personal health data linked to the EH Tracker.</td>
<td></td>
</tr>
<tr>
<td>• The anonymized data from the EH Tracker service may become valuable in research for generating new hypotheses about environmental health.</td>
<td></td>
</tr>
<tr>
<td>• People will be better aware of what they are susceptible to, thereby improving their quality of life.</td>
<td></td>
</tr>
<tr>
<td>• Checking pollution levels is now a daily routine for the general public.</td>
<td></td>
</tr>
<tr>
<td>• People are now better informed about their exposure time for the allergy-causing reagent while they commute to the work. The predicted discomfort score may help them decide where they work for the day: the office or their home.</td>
<td></td>
</tr>
<tr>
<td>• The doctor/patient relationship may improve, but also may be add another frustration as a result of increased communication. This may also apply to family ties and friendships. Teens and tweens also may not want their parent hovering over their every cough or sniff, etc.</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Box 3. Detailed evaluation plan for EH Tracker

In order to successfully design and maintain such a complex system for our diverse set of target users, our system will need to be evaluated continuously. This evaluation, the first of many, has several goals: 1) to define the characteristics of the EH Tracker design that will be measured, 2) to define methodology for measuring each EH Tracker characteristic, 3) to define administrative and participant evaluation procedures, and 4) to provide justifications for goals 1-3. It is our hope that successfully meeting each of these goals will guide a well-executed evaluation of the EH Tracker that helps our research team identify design problems, increase our understanding of user needs, and add to the collection of environmental health research artifacts that can be used by future research.

Design Characteristics

This evaluation targets the following design characteristics: ease of use, usefulness, action potential, and clarity. They were chosen for measurement based on their importance to users and to the overall system. It is imperative that the EH Tracker is easy to use. If users find the EH Tracker system difficult to use they may stop using it and miss out on its potential health benefits. In addition to being easy to use our system needs to provide useful information. Our overall goal, as stated in the introduction, is to enable our users to better understand how pollutants affect their health. Our assumption is that users who find this app useful will reap health benefits from such a detailed understanding of their environment. The underlying mechanism for health improvement in the EH Tracker system is behavioral change experienced through the provisioning of clear, actionable information. When a user is threatened by exposure to environmental pollutants, the EH Tracker needs to inform the user of the potential threat to their health and of steps they can take to avoid it. It is clear that each design characteristic plays a fundamental role in how the target users interact with the EH Tracker system. Thus, in addition to understanding these design characteristics, we also need to understand the design features that embody them.

In the context of this paper, a design feature is the manifestation of one or more design characteristics, and a design session is the evaluation of a single design feature. Thus, we can use a design session to simultaneously evaluate a single design feature and set of design characteristics. Design sessions will be created for the following design features: reminders, alerts, feedback entry, and the summary view. These features were chosen for two reasons: 1) each is comprised of measurable, target design characteristics, and 2) they are core features--the pillars of functionality that stand up the rest of the system—and therefore they are the most important features to evaluate. Table 1 defines each design feature and specifies the set of design characteristics that they will measure. Issues that are found during the evaluation of each feature will be cataloged and used to improve the feature after the session is complete. For example, if during the alert session users find that an alert icon is non-specific and confusing, evaluators will take note and use the user feedback to improve the design of the icon after the evaluation is complete. As a result, by validating design characteristics within core design features—the features that enable the rest of the system to function—we get the added benefit of improving the features that are the most important to the EH Tracker’s design.

Methods

Each design session will conduct a user based usability study followed by a user survey. The goal of each session is to better understand how users interact with and feel about the design features and characteristics under test. We chose to analyze our design features and characteristics in usability studies (instead of using another evaluation methodology) because evaluating how our users interact with our system in a controlled environment is an imperative insight to gain at this stage in the design process. At this point in our design development we have conducted extensive user research and developed detailed personas, scenarios, and prototypes, but we have not yet put our designs in front of users. Before we spend more time and energy further developing our designs we need to understand how our users interact with what we have. A user based usability study was chosen, as opposed to an expert based or automated usability study, because it's an affordable way for the evaluators to observe and record actual users of the EH Tracker system. An automated usability study would not allow us to capture real users using our system, and recruiting users in an expert based study would be difficult given that the EH Tracker system is in its infancy.
Supplementary Box 3. Continued

Following each usability study, a survey will be given to users to solicit feelings about the performed usability task. Surveys were chosen because they are a cheap yet effective way to capture detailed user feelings, which we will pair with our observations from the usability study to strengthen the results of our measurements.

Each design session, composed of a usability study followed by a survey, has the potential to provide a detailed understanding about specific user needs. Thus, it is imperative that each session is executed successfully. Below we define procedures for evaluators and participant to follow to increase the likelihood of a successful evaluation.

Procedure

Each evaluator will follow a set of procedures to conduct each design session, and each participant will be given a set of instructions to follow in order to successfully participate in the usability study and survey. At the beginning of each session evaluators will receive user consent to record the session on video. Once recording begins evaluators will provide participants with instructions on how to complete the usability study. These instructions will include how to properly use any devices, how to interpret the usability task description (see Supplementary Table 4), and other information specific to the usability task being tested. Users will be given adequate time to digest the instructions and read the scenario description. When the evaluators feel each user is adequately prepared to perform the usability task, the test will begin. Task completion time will be recorded for each user, then used to measure ease of use--each time will be compared to a predefined benchmark associated with the task. A time greater than the benchmark indicates the task took the user longer than expected and may be difficult to perform. A time less than the benchmark indicates the task took the user less time than expected and may be easy to perform. Right after usability tasks are complete, evaluators will hand out a short survey to each user containing questions about the task they just completed. Users will be asked to fill out the survey to the best of their ability, and will be given enough time to do so. When users are finished, evaluators will collect their surveys and catalog them for the research team to analyze when all design sessions are complete.

Supplementary Table 4. The usability tasks that users will be asked to carry out on each design feature, and the design characteristics that will be evaluated in each design feature.

<table>
<thead>
<tr>
<th>Design feature</th>
<th>Usability task</th>
<th>Design characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reminders: adds reminders for users to avoid pollution</td>
<td>Users will receive a fake email containing an upcoming event, for which they will be asked to add a reminder (as depicted in figure A.1).</td>
<td>Ease of use</td>
</tr>
<tr>
<td>Alerts: alerts users with actionable information about environmental health issues</td>
<td>Users will receive a fake popup alert. When they open the alert they will see fake, actionable environmental health information.</td>
<td>Ease of use, Usefulness, Action potential, Clarity</td>
</tr>
<tr>
<td>Feedback entry: solicits daily user feedback about their health</td>
<td>Users are given feedback (e.g. a good/neutral/bad feeling, a fake medication) to enter into the feedback entry system</td>
<td>Ease of use, Usefulness, Clarity</td>
</tr>
<tr>
<td>Summary view: summarizes the users environmental health history</td>
<td>Users will be asked to navigate to a summary page populated with fake information.</td>
<td>Ease of use, Usefulness, Clarity</td>
</tr>
</tbody>
</table>
Improving user engagement and insight through Contextualized Quantified Self

Tiffany J. Callahan, BA\textsuperscript{1,3*}, Tal Lorberbaum, BS\textsuperscript{2,3*}, Alexandre Yahi, MS\textsuperscript{2,3*}
\textsuperscript{1}University of Colorado Denver, Anschutz Medical Campus, Aurora, CO; \textsuperscript{2}Department of Biomedical Informatics, Columbia University, New York, NY; \textsuperscript{3}Plurl Health (http://www.plurlhealth.com)
* These authors contributed equally to the work; order is alphabetical.

Definition of the selected challenge related to the interaction between humans and data-analytic systems
The proliferation of smartphones and commoditization of health sensors (including step trackers, heart rate sensors, and blood pressure monitors) has enabled a process of self-digitization known as the “quantified self” (QS).\textsuperscript{1} The QS movement has caught the attention of the research community seeing in self-quantification systems (SQS)\textsuperscript{2} an opportunity to tackle P4 Medicine issues (i.e. predictive, preventive, personalized and participatory)\textsuperscript{3} by exploring the unique datasets generated by QS users.\textsuperscript{4,5} Companies such as Fitbit, Jawbone, Misfit, Withings, Apple, and others sell a range of reliable health and lifestyle tracking sensors\textsuperscript{6} and accompanying applications which have been met with sales exceeding 43 million units\textsuperscript{7} in 2015 and predicted to reach over 100 million by 2017.\textsuperscript{8} Coupled with health apps such as MyFitnessPal, Moves, and Lifesum, this ecosystem promises to combine lifestyle data and analytics to derive insights into health and suggestions for improvement.\textsuperscript{9}

However, a recent study highlights the striking reality that half of activity tracker owners stop using their devices, and one-third stop after only six months.\textsuperscript{10} The early-stage venture fund Rock Health recently suggested one of the potential reasons for this drop: wearable devices makers usually leave the use cases of their product to the imagination of their customers.\textsuperscript{9} Among the many challenges that need to be addressed, building software that identifies meaning in lifestyle data and can close the loop to provide actionable suggestions is critical. In addition, the market is very fragmented with apps and devices that are either single-purpose or so oriented towards the general market that they end up providing no insights at all. A population of core users called Quantified-selfers has overcome the multiple barriers discouraging many users\textsuperscript{11} but several pitfalls remain unaddressed. These include poor interface design (e.g. complicated data entry menus) and insufficient analytics in determining lifestyle patterns that are predictive of phenotypes important to the user.\textsuperscript{12} Apps often strive for entry granularity that is so high that it creates a burden on users, leading to less data collected than if the entry process were streamlined.\textsuperscript{13} Additionally, current data visualizations are limited to progress bars and individual plots for each dimension of data, making it challenging to observe trends in lifestyle data and opportunities for changes.\textsuperscript{14} Users often view their data in a vacuum with no context, making algorithmic suggestions for improvement unclear and further contributing to issues with user acceptance.\textsuperscript{15}

We propose an interactive user interface for Contextualized Quantified-Self (CQS). This interface puts quantitative and qualitative user data in the context of the user’s activities (e.g. meetings), environment (e.g. weather, pollution), and other users. By representing a user’s day with a unique visualization, it is possible to go beyond the “what to do” and break it up into actionable tasks to guide the user on “how to do it” when associated with a Learning Health System (LHS).\textsuperscript{16}

We present this visualization system’s ability to compare a day between users and between a user and a group using high blood pressure (HBP) as a case study.

Identified health problem
According to the American Heart Association, roughly 78 million American adults have high blood pressure (BP),\textsuperscript{17} only half of which have their high BP under control.\textsuperscript{18} What is more shocking is that in 2010 alone, more than 362,895 American adults died from high BP-related major cardiovascular disease,\textsuperscript{19} leading to an estimated cost of over $69.9 billion dollars.\textsuperscript{20} High BP can be diagnosed by a healthcare provider at a hospital, clinic or even at a local health fair. While the health risks of untreated high BP are concerning, it can be prevented and controlled. In fact, roughly 46,000 deaths each year could be prevented if 70% of adults with high BP received appropriate treatment.\textsuperscript{21} The American Heart Association made hypertension a primary focus of its 2014-2017 strategic plan and is hoping to improve cardiovascular health by 20% through reducing cardiovascular disease and stroke by 20% by 2010.\textsuperscript{22}
The Healthcare Effectiveness Data and Information Set (HEDIS) quality indicator is measured by determining, “the percentage of members 18 years of age and older who had a diagnosis of hypertension (HTN) and whose BP was adequately controlled during the measurement year using the following criteria”.

We believe that such an interactive system can have an impact on this type of quality indicators by displaying both recommended measurements and encouraging self-tracking. A versatile platform able to support a LHS for wellness is more likely to engage a patient to monitor his/her physiological parameters and could help hospitals reach out to their outpatient population.

Figure 1: Overview of Contextualized Quantified Self (CQS) visualization. (A) User interface displaying calendar (top), Day Ring (center), and Goals Panel (bottom). Each entry is added to the user’s timeline (Blue arc: sleep; red curves: steps with exercise intensity; pins: blood pressure measurement, meals, mood, water). The user’s day is surrounded by environmental information (e.g. weather). Personalized daily goals generated using a Learning Health System aggregating many users’ data are presented as a checklist, and the user can tap on the button to the right of each goal to see its justification. Remaining goals are listed as “ghost entries” on the timeline providing a suggested time for completion. Detail View in center displays additional information. (B) One vs. one comparison view allowing a user to compare him/herself against their activity on a previous day. Green bars indicate lines of agreement; red bars indicate discordant times and opportunities for improvement. (C) One vs. many comparison view. Inner circle displays time range during which team has completed activities. Textual summaries are also included for motivation.

Description of the proposed solution
We designed this interface to work with a Learning Health System (LHS) providing personalized goals leading to a long-term objective, and suggested time-dependent tasks on the visualization as guides toward goal completion. While the specifications of the LHS are beyond the scope of this design challenge, we will generate personalized daily goals for each user by comparing him or her against similar users to find lifestyles that are predictive of improved outcomes for the given objective. We then wish to display the results of these analytics in a clear and consolidated way. To do this, our interface is composed of three elements: the central Day Ring, the calendar at the top, and the Goals Panel at the bottom (Figure 1).
The central Day Ring contains several elements that allow the user to quickly see their data in context, view temporal relationships between their data, and see remaining goals for the day. A 24-hour clock contains the user’s timeline for a given day. Instead of treating each data type as its own isolated graph, we acknowledge and highlight the temporal relationships between different data types by displaying each actively or passively logged entry (including sleep, steps, meals, drinks, mood, and blood pressure measurements) sequentially on the timeline (Figure 1A). A user can tap on any entry to view more detail. Remaining goals from the LHS and suggested times for their completion are displayed on the user’s timeline as “ghost entries” that summarize the context-driven goals generated from many users back onto each individual user’s timeline.

The Day Ring supports three types of comparison: one versus one (e.g. one user against a previous day or one user versus another user), one vs. many (e.g. one user against the day of their team), and many vs. many (e.g. their team compared to the entire user population). In the center of the Day Ring is a Detail View that allows the user to see progress towards completing their goal, detailed measurement information, and environmental data such as the weather forecast.

Above the Day Ring is a calendar from which users can elect to display their previous days. At the bottom of the page, a panel displays the goals for that day. Unlike many QS apps where users must laboriously keep a diary, our interface represents goals using a checklist format that automatically updates when they have been reached. This rationale conveys more closure, and each goal can be expanded to get access to its algorithmic justification. The ability to expand goals to view the analytic basis for their inclusion will help relieve symptoms of the “black box” models and invite users to understand the rationale for suggested lifestyle changes.

Using our case study of monitoring blood pressure, our goals would take into account common recommendations for blood pressure measurement (e.g. consistent time of day for measurement, frequency of measurement, leaving adequate time after awakening and before meals and exercise to measure blood pressure) as well as diet and exercise suggestions. We will also be able to leverage environmental factors such as weather or pollen levels to suggest appropriate times for achieving exercise requirements for the day. Finally, lifestyle patterns identified using our LHS for both a user over time as well as multiple similar users would be summarized in the daily goals and ghost entries.

**Discussion of alternative solutions considered**

As alternatives we considered bar graph solutions as well as different shapes for the timeline. Bar graphs would entail having an individual chart for each data type collected. We also considered alternate shapes for the timeline as well as simplifications such as grouping times into morning, afternoon, and evening. Finally, we iterated on the proposed solution, beginning with just the timeline before adding the Detail View and environmental data. This was inspired by our realization that it was necessary to correlate a user’s activities with external factors such as environment.

**Discussion of the strengths and weakness of the chosen solution as compared to the alternatives**

The major strength of our design is that it can be adapted to a variety of screen sizes including smartphones and smart watches. The concentric design is adapted to concisely visualize multiple layers of information (e.g. physical activity, meals, sleep, mood, beverages, schedule, and environment). The weakness is that users may not be familiar seeing data visualized in this way compared to standard visualizations such as bar graphs. However, we believe that seeing many dimensions of data consolidated together will facilitate better understanding of the emergent trends our LHS finds in the data.

**Proposed implementation and dissemination plan**

As mentioned above, this interactive visualization interface is designed to directly interact with a LHS that, for a given objective, compares one user to many similar users to provide personalized daily goals broken up into time-dependent tasks displayed on the Day Ring. To this end, we are actively implementing the presented design in the form of an iOS app. Users will contribute both passively collected data (e.g. steps and sleep from wearable devices that communicate with HealthKit) and actively collected data (entered using a majorly optimized and simplified data entry menu also beyond the scope of this abstract). The data generated by users will be stored on a HIPAA-compliant server. This implementation is scalable since both the data storage and computing power necessary to support the LHS will grow with the user base. We will disseminate the app through Apple’s App Store.
Proposed evaluation plan
This interactive visualization interface aims to be objective-agnostic and therefore fulfill the expectations of a wide variety of profiles. We identified three different sub-populations for evaluation: users with no chronic medical conditions and not used to practicing self-tracking, core quantified-selfers already engaged in one or several SQS, and users with high blood pressure who require frequent monitoring. The recruitment of these users will go through an IRB protocol and they will sign a consent form to agree on sharing their data after de-identification.

Supplementary video demoing the visualization features can be found at:
https://drive.google.com/file/d/0B6sIoU_tAHoGyWdUNUp3cENHTWM/view?usp=sharing

References
20. Heidenreich PA, Trogdon JG, Khavjou OA, et al. On behalf of the American Heart Association Advocacy Coordinating Committee; Stroke Council; Council on Cardiovascular Radiology and Intervention; Council on
Clinical Cardiology; Council on Epidemiology and Prevention; Council on Arteriosclerosis, Thrombosis, and Vascular Biology; Council on Cardiopulmonary, Critical Care, Perioperative and Resuscitation; Council on Cardiovascular Nursing; Council on the Kidney in Cardiovascular Disease; Council on Cardiovascular Surgery and Anesthesia, and Interdisciplinary Council on Quality of Care and Outcomes Research Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. *Circulation*. 2011;123:933-44.


Take a Breather: Empowering Adherence & Patient Centered Research Through Interactive Data Visualization, Social Engagement, & Gamification in Patients with Sleep Apnea

Ashley Choi, Benjamin Cordier, Prerna Das, Jason Li
Division of Bioinformatics and Computational Biology
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR

Abstract

Sleep apnea is a chronic condition that affects over 100 million people worldwide, lowering their quality of life and productivity. Treatment of sleep apnea involves long-term management characterized by regular usage of a continuous positive airway pressure (CPAP) device and lifestyle modifications, such as reducing alcohol consumption and promoting weight loss through exercise and dietary adjustments. Together, these treatments have been shown to be highly effective in alleviating sleep apnea, however, adherence to CPAP therapy remains a primary obstacle to treatment success. To encourage adherence, we propose an interactive dashboard application that provides support and empowers patients to take ownership of their treatment regimen. Using this dashboard, patients will log their CPAP usage hours alongside biometric data (e.g. weight, blood pressure, heart rate) and sleep quality measures on a regular basis. From this effort, patients will gain the ability to visualize their health data and progress over time, in addition to being able to compare their advances to that of the overall cohort of participating patients. Patients will further be rewarded with socially visible awards for achieving preset health and community participation benchmarks. We expect that this application will encourage sleep apnea patients to participate in their own healthcare and take charge of their treatment plans. It is hoped that by fostering such positive dynamics, patients will experience improved health outcomes and quality of life.

Summary

With a geographically consistent prevalence of 4% in men and 2% in women, it is estimated that over 430 million people worldwide (and over 19 million in the United States) are affected by sleep apnea, with many remaining undiagnosed (1). Sleep apnea is characterized by the frequent occurrence of nighttime apneas, with an apnea defined as an episode of complete or near complete cessation of airflow due to upper respiratory collapse lasting at least 10 seconds (1). Perhaps of greatest concern, sleep apnea is associated with numerous serious health concerns such as hypertension, stroke, cardiovascular disease, diminished neurocognitive function, insulin resistance (1, 2), and sleep fragmentation, which has been found to increase the risk for motor vehicle accidents and a reduced quality of life (3).

As prevalence and awareness have increased, there has been continued focus on treatment options (4). Fortunately, effective treatments for sleep apnea do exist. Research has shown a linear relationship between the hours of CPAP use and improvements in quality of life and sleep (5). Similarly, it has been found that weight loss and abstinence from alcohol may reduce symptoms (1, 6). Yet, despite the established effectiveness of CPAP therapy, adherence remains a significant barrier to the treatment of this condition with compliance potentially being as low as 46% (7). Thus, a primary challenge in the management of sleep apnea and its associated risks are the adherence and long-term adoption of healthy behaviors to mitigate (and potentially eliminate) symptoms of the condition (8).

MyApnea.org, part of the Patient Centered Outcomes Research Institutes (PCORI) network, is focused on building a patient-centered community that empowers patients with sleep apnea to engage with sleep apnea research and facilitate adherence to their treatment plan. Despite being in the earlier stages of deployment, MyApnea has quickly gained traction with over 5,300 members and over 200 actively participating in MyApnea’s forum. Consequently, the desire to integrate features that engage users, encourage retention, and provide patients with a means of understanding their data in an intuitive fashion has grown significantly.

After connecting with the MyApnea team and assessing the user and site needs, we proposed the creation of an interactive dashboard with a data visualization centerpiece for MyApnea.org, with the aim of
fostering long-term engagement by patients with their health and treatment plan and encouraging user retention. While there are no universally agreed-upon feature sets or web design standards to achieve these objectives in the context of patient engagement, general principles and methodologies can be taken from popular, well-designed websites and web applications currently in existence. One such methodology that is widely employed by websites seeking to communicate novel ideas, complex datasets, or engage in data-driven storytelling is interactive visualization.

For the AMIA 2015 Student Design Challenge, our aim was to build a prototype of a tool capable of achieving these objectives. During our initial ideation, we came to feel that the integration of two additional methodologies would increase the potential of the tool achieving its objectives: social dynamics and gamification. Specifically, it was felt that by 1) providing users with the ability to compare personal progress with site-wide progress, a competition for better health may ensue and 2) allowing users to set their own personal health improvement goals, in addition to attaining preset, socially visible, shareable achievements, a sense of accomplishment and satisfaction—ultimately encouraging site engagement, retention, and perhaps treatment adherence—could be facilitated. Together, these features would be presented in an intuitive and appealing dashboard using established and readily comprehensible data visualization techniques.

The Challenge

MyApnea.org is part of a patient-powered research network that brings patients, doctors and researchers together to improve the health of patients with sleep apnea. There are currently about 5,300 registered members in every state of the U.S and in 41 countries worldwide, with an enrollment goal for the project of 50,000 patients. Within the site, there are information sections about sleep apnea and the various treatments, a community forum where users can interact with one another, and a Rank the Research tool where patients can influence the direction of sleep apnea research by voting for topics that they find to be relevant to their experience. Additionally, there are 11 research surveys that collect a wide variety of patient data regarding family history, health conditions, sleep patterns, sleep quality, and quality of life. Each survey has about 10 questions and, upon completion, shows the user their answers in the context of individual answers across the community using a pie chart visualization. At the time of writing, over 1,000 users have completed 36,000 surveys. To assess the needs of MyApnea, we consulted with the researchers, software developers, and a leader of the MyApnea participating patient cohort. Across these stakeholder groups, a commonly cited need was that of a feature to promote website retention; a critical component to achieve high-frequency, large cohort data collection and conduct an anticipated longitudinal study.

At the time of writing, the MyApnea team was working on implementing a semi-annual survey to collect data for the aforementioned longitudinal study, however, it was felt that a more frequent mode of engagement would be advantageous by allowing for more opportunities to collect data from participating sleep apnea patients. From such increased participation, a more complete record of each patient’s sleep apnea could be created and, perhaps, through research, a more granular understanding of the data could be yielded. This, along with the possible added benefit of increased nightly CPAP usage by the engaged patients. Thus, broadly, the challenge: Create a feature set capable of increasing patient engagement and data contribution and, if and when engagement does increase, have the feature set be capable of leveraging this engagement to create a positive feedback, further encouraging patient participation.

Evaluation of Possible Solutions

A number of methods were considered to increase user retention, including 1) providing new and regularly updated content, 2) building a community, 3) making a game-like experience, and 4) providing monetary incentives or similar benefits.

New and updated content can come in the form of either curated articles, as on a news site, or user-generated content (updates, photos, etc.) on a social media site. Given that news updates specific to sleep apnea may be fairly sparse, they represent an unreliable source of updated content. Consequently, the primary source of updated content would likely be user-generated, through each user’s own surveys, their regularly submitted health data, and similar site-wide data.

Online community building can come in the form of user forums, comments sections, and other forms of social media. Size often plays a significant role in the vibrancy of any community. Social sharing is a fantastic way of expanding a community and publicizing a website’s mission. By allowing users to share visualizations of their progress and achievement medals through social media websites such as Face-
book, Twitter, Instagram, etc. pre-existing personal connections within the MyApnea community are reinforced while the website is shared family and friends who may have similar interests or suffer from sleep apnea themselves.

Gamification is a serious business. Smartphone games such as Candy Crush Saga account for an estimated $30 billion in revenue worldwide in 2015 (9). According to a survey conducted in late 2014, people spent over 2 hours a day playing games on mobile devices, up from 1 hour and 20 minutes in 2012 (10). Often, there are no tangible rewards for playing these games, but the psychological boost from achieving predetermined goals or gaining the next level is addictive. Giving out rewards like badges and stars for completing surveys, filling out data forms, and CPAP use may be effective in bringing users back regularly. In addition, the use of progress bars to visualize distance to a future achievement has been shown to encourage user engagement.

Even with the inclusion of such modern retention methodologies, ultimately, MyApnea must benefit its users sleep apnea and overall health. The proposed visualization, alongside the stated methodologies, will help the patients on MyApnea improve their symptoms and quality of life by improving their understanding around the connection between BMI, alcohol, CPAP usage, and their sleep apnea while, hopefully, empowering them to adhere to their treatment plan.

**Conceptual Implementation of Solution**

Our proposal is for a web application to visualize user data on MyApnea.org. The application is divided into three main modules: Visualization (‘Progress’), User Input (‘Update’), and Gamification (‘Medals’).

**Visualization Module**

The first module is for the user-selected visualization; comprised of two time series charts, with the upper time series including biometric and CPAP data sources the lower time series including sleep quality measures. Data filters will be used to allow the user to visualize combinations of these data sources. The data sources that will be displayed include BMI, systolic blood pressure, diastolic blood pressure, heart rate, perceived daytime sleepiness, number of nighttime arousals, and perceived sleep quality. It is expected that, with increased CPAP usage, sleep quality will increase and daytime sleepiness will decrease. Additionally, there should also be an approximately inverse relationship between BMI and sleep quality. Aside the data filters, a data toggle will allow the user to select between only their data or a comparative visualization of their data against site-wide averages. From this, a user will be able to assess how they are doing in respect to all the users across the site, encouraging a personal goal of ‘beating the averages’. Finally, panning and zooming of the data will allow the user to hone in on points of interest, enabling a more granular examination of patterns in the data and, more generally, their progress.

**User Input Module:**

The middle module will be for users to enter their ‘daily’ data. It has input fields for weight, CPAP usage hours, systolic blood pressure (SBP), diastolic blood pressure (DBP), heart rate, perceived daytime sleepiness (1, i.e. very poor – 10, i.e. excellent), perceived sleep quality (1, i.e. very poor – 10, i.e. excellent), and the number of nighttime arousals. BMI is computed from the height provided in the surveys and the weight provided each time a user submits data. For best utility, the user should enter data daily. In the future implementation of this application, a user’s CPAP may be able to upload data and fill in several of the fields directly via a data card or USB cable.

**Gamification Module**

The bottom module will show the user’s achievements through points and star badges. Achievements will be given for the completion of milestones relating to community participation on MyApnea.org or for healthy behaviors, as indicated through data submitted through the input module. Each achievement will give the user a certain star badge, notifying them with a small pop-up notification. There will be progress bars below the stars to indicate progress toward the next milestone. For instance, responding to another user’s forum post earns a bronze badge for the category “Forum Friend”, while posting 10 times...
earns a silver badge. Similarly, to become a earn a gold star as a “CPAP Warrior”, one must submit CPAP usage data every day for 6 months. Other achievements include “Survey Samurai” for survey completion, “Data Dignitary” for daily data submission streaks, and “Thought Leader” for receiving up-votes for a research question posted using the Rank the Research tool, with many other possible titles to encourage user goals. These badges will be visible to other users on the forums, and will be able to be shared via a variety of social media platforms: when one earns a badge, a dialogue will prompt the user to share the achievement via Facebook, Twitter, or Instagram.

Survey Module

It should also be noted that we are currently engaged in an ongoing exploration of other visualization possibilities for the survey data. It is likely that if such a visualization is created, it will be done with a new Survey module.

Software Architecture

The underlying implementation will be a modified model-view-controller architecture (Figure 1.) Data entered will be posted to the server through a REST API using SSL encryption, ultimately updating a database of all the user data. The visualization module will be rendered using data requested from the server via the REST API, with the server making the appropriate query to the database. Currently, MyApnea.org is built on an Apache server using the Ruby on Rails framework, coupled with a PostgreSQL database. The front-end of the application will be implemented in MyApnea.org’s current responsive CSS framework.

Prototype Implementation and Evaluation Plan

The current prototype is a simplified, standalone version of the conceptual dashboard application. It is an illustrative example of the dashboard concept articulated above with the purpose of facilitating communication of the dashboard concept to the MyApnea team and participating patient cohort. At the time of writing, we are engaged in discussions with the MyApnea team regarding the possibility of integrating the dashboard concept into the MyApnea.org web application. Here is a link to the current iteration of the prototype.

Visualization Module

The primary difference between the prototype and the conceptual implementation described above, in regards to the data visualization module, is the absence of data source toggling; social data is not available in the prototype. In addition, the prototype is currently operating on mock data generated within the application. This mock data is generated on a ‘per patient’ basis and illustrates relationships within the data commonly seen in the real-world: systolic blood pressure is higher than diastolic blood pressure, BMI is a function of a person’s weight and height, increased CPAP usage results generally (with some noise) improves sleep quality, decreases daytime sleepiness, and decreases the number of nighttime arousals.

User Input Module

The prototype includes a simple UI mockup of the user input module with a web form containing data input fields and a submission button. With the current implementation, the user can fill out the data input fields of the form and click the submission button, however, no data will be sent to the server or saved to the database. Front-end validation of the input data has yet to be implemented, however, such validation will be implemented on any deployment of this module on the MyApnea.org website.

Gamification Module

While the user interface for this module is mostly functional, it is not data driven. Instead, alterations to the underlying code have been made to illustrate how the conceptual implementation might look once a user has achieved a number of medals. The social media component has yet to be implemented, too.
Survey Module

For the prototype, no survey module was developed. Exploration of possible uses for the survey data in the context of user engagement is ongoing.

Additional Technical Details

The server environment has been written in JavaScript using Node.js and the Express server framework for rapid prototyping. The software has been deployed using Heroku, a cloud-based platform-as-a-service (PaaS) host. The prototype is not responsive to browser and device screen dimensions, however, a live iteration would involve leveraging the pre-existing responsive CSS framework on MyApnea.org.

Evaluation

Evaluation will focus on garnering feedback from the stakeholders working on the MyApnea.org website. Feedback from the researchers and software developers building MyApnea has been, and will continue to be, critical to the dashboards success and upcoming integration with MyApnea.org, however, perhaps more importantly, a dialogue with the MyApnea patient cohort has begun regarding the prototypes usability, objectives, and the addition of features that users would like to see in upcoming iterations. Given the modular and extensible layout of the dashboard interface and underlying software architecture, it is possible that, from this feedback, this engagement may grow to include new modules that further encourage user engagement, retention, and perhaps an improved quality of life for its users.

Conclusion and Further Work

The proposed interactive web application will increase patient engagement, adherence, and retention. It will accomplish this through providing patients with a way to view their progress and compare themselves to site-wide averages, rewarding them for good health behaviors and adherence to their treatment plans using achievement badges. At the time of writing, a live, full-featured implementation has yet to be deployed to patients and undergo feedback-based development. Such iterations will likely be necessary to develop a tool that truly accomplished the outlined goals.

Currently, the conceptual dashboard only incorporates survey and patient-entered data sources, however, in the future, the MyApnea team anticipates collecting quantitative data through each user’s CPAP directly, which would likely prove an invaluable data source. As patients acclimate to the dashboard, they may appreciate having a means of easily uploading their sleep and CPAP usage data. Such an ability would lessen the time commitment when entering data and provide a quantitative, less bias confirmation that their behaviors are indeed improving their sleep health. Furthermore, the dashboard concept illustrated here has the potential to be expanded and adapted to data relevant to other health conditions, thereby making it a tool for any health website that wishes to increase patient engagement.
Figures

Figure 1. A representation of the dashboard concept’s underlying architecture. When the application is loaded, an initial request is made to the server using a custom REST API to retrieve the user data that drives the visualization and gamification modules. This data is created through user submissions using the Input module. After the user fills out the Input module’s form and clicks the submit button, the data is sent to the server layer. The data is then validated and processed into a record object that is representative of the database schema. The record object is then submitted to the database. It is from these records that data is pulled for the Visualization and Gamification modules.
References


10. Rad C. Mobile game play time up 50 percent since 2012. IGN. 2015.
I-SMILE: Similarity based Just-in-time Recommendation System for Public Health
Edward Choi, MS1, Jina Dcruz, PhD2, Sizhe Lin, BS1,
Kelly Ryder, MPH3, Aashu Singh, BA1, Hang Su, BA1
1Georgia Institute of Technology, Atlanta, GA
2Health Informatics Fellow at Centers for Disease Control and Prevention, Atlanta, GA
3Association of Schools and Programs of Public Health (ASPPH), Washington, DC

1. Problem Definition
Public health and informatics practitioners are often required to apply and conduct a research that involves knowledge outside their expertise. Epidemiologists, for example, could be assigned a task to analyze the seasonal trend of influenza by studying Electronic Health Records(EHR), which requires certain information processing knowledge such as data structure, programming language(e.g. Python), and database management. There are various ways to acquire new knowledge: traditional methods such as taking lectures or reading books and papers, taking advantage of the information technology by searching the Web. Recently, thanks to the high-speed Internet, massive online open courses (MOOC) such as Coursera has gained much popularity. Such methods, however, assume that the user knows what to look for, as they all require a certain searching phase (e.g. need to search for the appropriate lecture, need to use the appropriate search keyword).

But what if the user does not know what to look for? It is very likely that the epidemiologist from the above example is not familiar with the terms such as Natural Language Processing, Python or NoSQL Database Management. Researchers from the medical/health field often face this problem these days, especially now that computing, statistical and analytic methods are actively being exploited. Effectively addressing this knowledge gap will help many interdisciplinary studies to be conducted with a smoother learning curve. This problem can be significantly alleviated by involving human interaction in the search process. For example, by describing his/her given task, the researcher can guide the machine to look for relevant learning materials. Also, the machine can take feedbacks from the user so that it can improve its performance as it repeats the search. Teaching relations between various concepts (or knowledge) to the machine can also help the user to navigate through the recommended learning materials, as opposed to scrolling through the list of results.

In this paper, we present I-SMILE, a project lead by the students of Georgia Institute of Technology in collaboration with the Centers for Disease Control and Prevention (CDC). I-SMILE is a similarity-based recommendation system that understands the task described by the user, and recommends relevant books, papers, online courses, and even human experts. The system utilizes the project description and also considers the expertise and learning profile of the user, what he/she searched for in the past, so as to better recommend the most relevant learning material to the user. The user can also search through concept taxonomy such as MeSH Tree to get a high-level understanding of relevant concepts and navigate through the recommended learning materials.

![Figure 1. System Architecture](image-url)
2. System Description

Figure 1 depicts the overall architecture of I-SMILE, grouped by high-level components. We will go over each component to describe the functionality and the design choices.

Data Collection & Indexing

I-SMILE takes advantage of various forms of learning materials that can be collected from the Internet. We have collected the course list and course descriptions from Coursera and Udacity, two popular MOOCs. Book descriptions were collected from Google Books, where MeSH terms were used as search keywords. Abstracts of papers and journals were collected from PubMed and PubMed central, again using the MeSH terms as search keywords. Additional educational materials and lectures were provided by the CDC. All collected learning materials are then processed using basic natural language processing (NLP) techniques such as lemmatization, and indexed into a traditional vector-space model information retrieval (IR) system.

User Profile

Users must create a user account in order to use I-SMILE service. Figure 2 (a) is the main page of the current version of I-SMILE. At the top is the Sign Up link, where the new user can create an account.

When first creating their accounts, users can also enter their skill sets and the level of proficiency for each skill. This information will be used in combination with the user feedback, which will be described later, to search for learning materials that specifically suit each user’s need. Specifically, the user profile and the user feedback is leveraged by collaborative filtering, a well-known recommendation algorithm. We plan to implement collaborative filtering module in the next development phase. The collaborative filtering module will play an important role in scoring the learning materials in the searching phase.
**Task Description Analysis (Problem Narrative Search)**

One of the key features of I-SMILE that involves human interaction is its ability to support narrative search. In the text box in Figure 2 (a), the user can provide a text that describes the task he/she needs to fulfill, and the machine will analyze it to extract important keywords. I-SMILE uses a part-of-speech (POS) tagger and a dependency parser from Stanford NLP group’s CoreNLP package. Given a text, the system analyzes the text to extract only noun phrases. Then, using a list of stop words, it removes/transforms the extracted noun phrases to produce only the most relevant, distinguished noun phrases, which are in turn used as search keywords. An example can be seen in Figure 2 (b), where extracted words are highlighted and added to the list on the right. The user can check the list and remove/add new keywords. We are currently adding a feature to use MeSH terms and document frequency of each word to improve the performance of keyword extraction.

Additionally, the system provides a list of MeSH terms that are most similar to each extracted noun phrase in order to help the user decide if he/she needs more search keywords. In Figure 2 (c), the system is suggesting MeSH terms that are most relevant to the word **natural**. You can see that **Natural Language Processing** placed on the top of the list.

**Social Networking (People and Resource Finder)**

I-SMILE differs from other recommendation/search engine in that it can find users who are working with, or worked with similar tasks. The motivation behind this feature is to promote collaboration and knowledge sharing between fellows working on similar projects. When the user performs a narrative search, the system stores the analyzed result of the user-given task description in its database. Then, when another user performs a narrative search, I-SMILE compares the analyzed result of the newly given task description with its stored results. The comparison is based on the Jaccard coefficient of the nouns extracted from task descriptions. If there is a certain level of similarity, the system generates a list of users who searched for that similar task description in the past. This feature involves exposing personal information such as user’s affiliation and user’s email address. Therefore, when first creating a user account, the user can opt-out to become unsearchable by other users.

**Search & Combine**

Using the phrases and words extracted from the Task Description Analysis phase, I-SMILE retrieves the relevant learning materials from its database based on the standard vector space model. The list of phrases and words are aggregated into an “or” search operation. In the next development phase, I-SMILE will also retrieve the recommended learning materials from the collaborative filtering module. Combining the results from the both modules, I-SMILE will return to the user the learning materials that are both relevant to the task description and suited to each user.

**Result Presentation & Navigation**

The combined results are by default presented in a traditional ranked list fashion, as can be seen in Figure 2 (d). The user can use the filter features on the left of the result page to selectively view the learning materials from specific sources. The list of users who searched for similar task descriptions is listed on the right side of the result page. By clicking the username, the current user can check the target user’s contact information and the task descriptions he/she searched for in the past.

If the user is not familiar with what he/she is searching for, then simply listing the retrieved learning materials in a relevancy order might not be very effective. Therefore, we plan to update I-SMILE so that it can import external concept taxonomy such as MeSH Tree. Once the system is given the taxonomy, it will link the documents in its database to each node of the taxonomy. After the search is performed and the system returns the learning materials to the user, he/she will be able to expand the taxonomy from left side of the search result page. Nodes that are relevant to the current search will be highlighted. By looking at the parent and child nodes of the highlighted nodes, the user can get an overview of the knowledge structure. The user will also be able select or de-select the nodes of the taxonomy, and the search result will change accordingly. Each time the user highlights or de-highlights a node, the previous search keywords will be modified, and the new search will be performed to match the current state of the taxonomy tree.

**User Feedback**

The user can give explicit feedback to the system by using the star-rating system, as can be seen in Figure 2 (d). If he/she finds a certain learning material useful/useless, he/she can give a rating to that specific learning material. I-
SMILE, however, can also take advantage of implicit user feedback, namely the click-through data. Although not as obvious as the star-rating system, a user clicking a certain learning material is also a form of feedback. I-SMILE stores this click-through information and uses this in combination with the user profile to perform recommendation based on collaborative filtering. Explicit feedback based on the star-rating system is utilized as a measure of evaluation, which will be discussed later. Click-through data can also be used as a measure for evaluation as well.

3. Comparison with Alternative Solutions

I-SMILE provides an interactive experience for the user. The important features of this system can be better understood in comparison with currently available alternatives for learning new knowledge.

Google Search

Google returns the search result as a list of links to articles/pictures/videos. Google search expect short phrases as query and does not work with long queries such as project description as in our case. Google search does not differentiate the sources of the results since everything is presented in a single ranked list.

I-SMILE periodically collects high quality learning materials from chosen sources, and makes them available to the user by combining both content-based vector space model and personalized collaborative filtering. This enables I-SMILE to provide new learning materials readily to users while considering the characteristic of each user.

The most important distinction between I-SMILE and Google search, however, is that the user can perform the search based on his/her task description, which does not force the user to be familiar with what he/she wants to learn.

MOOC Courses

Another way of learning new knowledge is to search for courses on MOOC sites such as Coursera or Udacity. A great feature of this learning method is that in some cases, the user might be able to network with his peers who are taking similar courses and help each other learn through collaboration, which motivated us to implement a networking feature of I-SMILE as described above.

Searching for online courses, however, also requires the user to know the proper search keywords in advance. Furthermore, typical MOOC websites only provide lecture videos, which are not always the best method to learn. Videos cannot be searched by its content, forcing the user to search lectures by their descriptions. I-SMILE provides multiple forms of high quality learning materials, including videos from two popular MOOC sites. Such availability enables the user to choose the learning medium with which he/she is comfortable.

Limitation

I-SMILE requires advanced NLP techniques to extract important keywords from the task description. Currently, the system uses Stanford CoreNLP package, which provides the state-of-the-art NLP modules such as a POS tagger and a dependency parser. These are, however, developed for general use cases, and therefore are vulnerable to very domain-specific text. For example, if the user-provided task description contains a lot of domain-specific acronyms or jargons, the NLP module will not be able to recognize them properly, leading to suboptimal results. The user can still modify the extracted keywords, but it will not be very effective without the domain knowledge.

This problem can be resolved if there are available training data to re-train the NLP modules for the target domain. Such training data, however, require considerable amount of time and money to build. In future, we plan to explore the recent neural network word representation models, which can be trained in an unsupervised fashion. Incorporating the word representations as features to the traditional NLP methods has proven to be quite effective. Word representations can also be utilized in other ways such query expansion or automatic construction of concept taxonomy.

4. Implementation and Dissemination

I-SMILE is an active ongoing project, being developed and tested on Amazon Web Service (AWS) cloud computing environment. Our system is exposed as a web application and the user can access the service by a web browser, without having to install any additional software. The back-end of the system is implemented mostly in Python, and the front-end in standard HTML and Javascript for web browser compatibility. We chose MongoDB to store the raw descriptions of the learning materials, and ElasticSearch to serve as the standard IR system. Both MongoDB and ElasticSearch are designed to be highly scalable and lightweight, which makes it very easy to expand I-SMILE to hold enormous amount of learning materials and serve the user in real-time. Furthermore, thanks to the flexible and
efficient cloud-based environment, large-scale deployment is as simple as adding more machines to the current cluster.

In the future, we plan to create VM images of the machine instances in our deployment environment for services providers who need a customized system. A new I-SMILE instance can be easily launched using the images we provide. Within minutes, they will have their own system running in their own infrastructure. We believe this feature will facilitate the adoption of I-SMILE. For example, organizations that manage sensitive data, such as research institutes will be able to safely operate the system by launching the image on their infrastructure and importing their internal documents and user profiles into the system without exposure to external users.

5. Evaluation Strategy

By regarding I-SMILE as a vertical search engine, conventional evaluation metrics for IR systems can be applied to evaluate I-SMILE. We plan to evaluate the system based-on both explicit and implicit user feedback with 50 informatics fellow from CDC (note that two fellows participated in this project from the start)

Evaluation with explicit feedback

We can construct an evaluation dataset by collecting the star-rating feedback by the user. For each recommended document \( d \), we associate it with the input task description \( q \) and the rating score \( s \) from user \( u \). Initially, we will ask the CDC fellows to carefully evaluate top 30 results for each of the 20 task descriptions and at least 3 different users should have a consensus in terms of result rating to give us reliable data for evaluation. Then we could measure the performance of our system with below metrics.

1. Precision measures the percentage of relevant document in a ranked list. We are interested in precision@1, precision@5, prediction@10, where precision@k is defined as below.

\[
\text{precision}@k = \frac{\#\text{relevant document in top-}k}{k}
\]

2. Discounted cumulative gain will take the score into consideration. The \( s_i \) in below equation is the rating of the material at position \( i \). Clearly, DCG gives more importance to higher-ranked documents. This is important because the user is unlikely to scroll down or even move to the next page to find interesting learning material if they can’t find one in the top ranks.

\[
\text{DCG}@k = s_1 + \sum_{i=2}^{k} \frac{s_i}{\log_2 i}
\]

Evaluation with implicit feedback

Though explicit feedbacks provide more accurate and reliable data for evaluation, collecting them requires considerable resources, which makes them inappropriate for a constantly changing project such as I-SMILE. We also plan to evaluate the system with implicit feedbacks, namely click-through data. The assumption behind this evaluation is that higher-ranked documents that are not clicked are less relevant than lower-ranked ones that are clicked. For example, given a list of learning materials \( (d_1, d_2, d_3, d_4) \), if only \( d_3 \) is clicked, it is believed that \( d_3 \) is more relevant than \( d_1, d_2 \). We maintain a database to record the click information of the user to leverage them in the future as a dataset for evaluation.

6. Conclusion

In this paper, we presented I-SMILE, a novel recommendation system that facilitates learning new knowledge for users who are unfamiliar with the target domain. I-SMILE employs various strategies based on NLP techniques to provide description-based search, social networking, customized search results, and taxonomy based result navigation. In the future, we plan to apply more advanced method such Deep Learning to improve the system’s capability to understand domain-specific task descriptions.

References


Learning from the Data: Exploring a Hepatocellular Carcinoma Registry Using Visual Analytics to Improve Multidisciplinary Clinical Decision-Making

Michelle Hribar PhD, L. Nelson Sanchez-Pinto MD, Kate Fultz Hollis MS, Gene Ren, Deborah Woodcock MBA
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR

Summary

Collaboration among cancer care providers -- often done in multidisciplinary conferences -- can improve patient care. However, managing the data associated with these conferences is difficult and there is often limited access to national level data for treatment decisions (e.g., national cancer registries). We propose a novel method for developing data visualizations and visual analytic tools to explore hepatocellular carcinoma (HCC) data from a national registry and implement them in the clinical workflow of multidisciplinary conferences. The HCC Data Visualization Repository and Clinical Tool will allow clinicians to extract complex information from HCC data using custom-designed visualizations. This solution would be extensible to other cancers and disease groups, and could significantly improve patient care. An interactive prototype of the repository may be accessed at http://hccviz.azurewebsites.net/.

Background

Problem definition. Since the early 1990s, cancer care became more complex due to increased specialization among disciplines and more sophisticated treatments, as well as greater access to clinical practice data in electronic health records (EHRs) and other biomedical “Big Data.” As a result, patient care is no longer provided by a single specialist; rather it is delivered by multidisciplinary teams (MDTs) working in conferences, commonly referred to as tumor boards. A recent systematic review of oncology MDTs and patient outcomes suggests that providing multidisciplinary care improved patient outcomes with regard to diagnosis and/or treatment planning, survival, patient satisfaction, and clinician satisfaction¹.

Tumor boards take many forms (e.g., in-person or virtual round tables, clinics) and may include oncologists, radiologists, surgeons, pathologists, nurses, social workers, dieticians, physiotherapists, and occupational therapists. Their goal is to not only improve patient care through collaborative decision-making, but also foster learning within the team that may lead to future care improvements. In the latter sense, tumor boards have the potential to function as “rapid-learning health care” laboratories within their own institutions. A rapid-learning health system is one in which EHRs, big data (data with high variety, volume, and velocity), computers, and learning networks are used to accelerate evidence-based medicine. Using this data for rapid learning will require tools including mathematical approaches, software systems, visualization platforms, and education². Despite their aspiration to learn rapidly, tumor boards have limited resources to make the rapid-learning process a working reality.

Toward rapid learning in cancer care. At Oregon Health & Science University (OHSU), the team caring for patients with hepatocellular carcinoma (HCC) meets weekly to formulate and refine treatment plans for 12-15 patients. A nurse coordinator prepares patient summaries, which are presented in paper form. As each case is orally presented at the multidisciplinary liver tumor conference, a team member navigates through the patient’s electronic medical record (EMR), projecting additional information (e.g., imaging studies) on the conference room screen (Supplemental Materials, Exhibits 3 and 4) (Observation at Multidisciplinary Liver Tumor Conference. 2015 Apr 14, 21, 28).

The OHSU liver tumor conference encounters several challenges. The first is inefficiency: much extra time is spent preparing summaries before the conference and hunting for relevant data within the patient’s EMR during the conference. The second is temporality: the difficulty in analyzing patient data over time (as a patient may be discussed on multiple occasions as their disease progresses). The third is limited access to relevant and potentially helpful data: the lack of a tool that enables HCC MDT members to compare their patient data with similar patients in a way that supports evidence-based care.
In response to these challenges, the OHSU liver tumor board is developing the Multidisciplinary Bioinformatics Data (MDBD) Tool, a software application to capture and store the rich knowledge generated from their tumor board meetings in a more structured way. Other institutions are collaborating with OHSU in a data-sharing network to develop a national HCC database for clinical care and rapid healthcare learning. The MDBD development team has not yet determined how providers will access the national level data; we propose that interactive data visualizations have the potential to present information so that trends and insights can be quickly identified.

Visualizing complex clinical data is difficult. In their systematic review of data visualizations of EHR data, West et al. found few EHR visualization techniques before 2014 that could process huge amounts of clinical and other data. “Although most studies recognize the importance of the growing amount of clinical data, we found few innovative EHR visualization techniques that lend themselves to the large amount of data available electronically.”

The AMIA challenge. Our team therefore chose to focus on the design of a platform for the development of data visualizations and visual analytic tools to interact with complex cancer patient data. Using the OHSU liver tumor board as our test case, we propose an application that will allow providers to explore the MDBD database with intuitive, powerful data visualizations developed in collaboration with HCC providers. This application will allow clinicians to identify national cohorts of patients with phenotypes (and eventually genotypes) similar to the patients they are treating, and to use the experience of many other medical teams to inform the decisions they are making; in short, to support an evidence-based rapid-learning system for HCC care.

**Solution Design and Development**

**Design process.** To determine the best approach for clinicians to interact with a national HCC database, we performed a needs assessment through interviews with members of the MDT liver tumor board and stakeholders of the Multidisciplinary Bioinformatics Data (MDBD) tool currently in development. Dr. Willscott Naugler, a gastroenterologist and lead Principal Investigator of the MDBD project, provided us with the workflow for the conference (Supplemental materials, Exhibit 3) as well as the vision for the MDBD project. The goal of the MDBD project is to improve the MDT conference workflow by streamlining data collection and presentation and ensuring data is structured and accessible to all MDT members. Further, all data collected by the tool will be de-identified and sent to a national HCC database. Participating institutions will have access to the detailed data about their own patients as well as data from other HCC patients in the national database.

Discussions with the head of MDBD development (Justin Ramsdill) provided us with the data dictionary for the tool. In addition, we discussed the needs for data interaction among members of the OHSU liver tumor board (Supplemental Materials, Exhibit 5). Since interacting with national HCC data isn’t currently possible, the liver tumor board members were not sure of the best approach. However, all expressed interest in seeing data about different strategies and sequences of treatment, particularly for patients clinically similar to those discussed at conference. They were interested in seeing the prevalence of certain treatment protocols as well as the outcomes for the protocols. They weren’t sure of the best way to view this data, but they wanted to be able to compare specific treatments and protocols quickly.

**Solution description.** Our proposed solution is a data visualization and visual analytics platform. This platform will have two main components: 1) an online visualization development and evaluation repository and forum; and 2) a data visualization and visual analytics tool integrated with the MDBD tool to aid with clinical decision-making during MDT tumor board meetings. This solution will take advantage of the human ability for pattern recognition and the capacity to easily extract relevant information from visualizations in order to improve care based on complex clinical data.

The online visualization development and evaluation repository is a key component of our proposed solution. The repository will become a forum for data visualization experts, informaticians, data scientists, and clinicians where visualizations and visual analytics solutions are conceptualized, developed, and improved through an iterative process. The workflow for this platform starts with a need or a question, such as: “What is the laboratory profile of patients who respond to a specific HCC therapy a month after their treatment?” Through a requirements-gathering process, clinicians’ needs are further refined and additional visualizations and visual analytics solutions are developed using the de-identified HCC sample data.

These visualizations are posted in the online repository, where feedback is gathered from users (both clinicians and non-clinicians) and used to make further quality improvements to the visual solutions. High-interest visualizations will be formally evaluated with a select group of HCC clinicians and beta-tested by all users of the MDBD tool. This online visualization development platform will provide a constructive solution to the uncertainty of the providers’
needs for data interaction. For illustration purposes, a prototype visualization repository along with three prototype visualizations have been developed and posted online (http://hccviz.azurewebsites.net/) (Supplemental Materials, Exhibit 1).

The clinical data visualization and visual analytics tool is the second component of the proposed solution and will be integrated with the MDBD application. The visualization and visual analytics tool will act as the interface between clinicians in the MDT meetings and the national HCC database. This application will allow clinicians to explore therapeutic and outcomes data of HCC patients on a national level using data visualization and visual analytics tools previously validated through the online repository as described above. Clinicians will be able to use this program as an additional tool to inform the decisions they make about their patients during the MDT tumor board meetings.

**Alternative Solutions**

Currently the OHSU Liver Tumor Board MDT reviews and discusses each patient’s case to come to a consensus about staging (liver and cancer) and a treatment plan. Staging and treatment decisions are currently informed by the Barcelona Clinic Liver Cancer (BCLC) Staging System (Supplemental Materials, Exhibit 2). The BCLC staging system provides rough guidelines, but not exact recommendations, particularly for non-surgical patients. Consequently, there is considerable variation in care, especially across institutions. Having access to national data about treatments and outcomes would help inform HCC treatment decisions; visualizing this data would help clinicians quickly gain insight.

There are few visual analytics tools currently used for comparing individual patient data with national databases. Searches on PubMed using terms like "Registries"[MeSH] AND visualization[All Fields] and "Medical Record Linkage/methods"[MAJR] AND visualization[All Fields] brought little recent references relevant to the use of data visualization with either internal registry data or national databases like SEER (Surveillance, Epidemiology and End Results). Maier et al. used the U.S. SEER cancer registry dataset with a digital breast cancer tumor board at a German university hospital to determine the dataset’s utility in the clinical settings. They noted that few studies in the literature specifically linked patient data with a cancer registry database. Other large databases with cancer data include CancerLinQ, the American Society of Clinical Oncology (ASCO) health information technology initiative whose first elements will debut by early 2015. Currently, though, there is no significant use of CancerLinQ in the clinical setting.

The MDT could interact with the national HCC database directly by using pre-existing business intelligence software such as Tableau (www.tableau.com) or QlikView (www.qlik.com). However, choosing a program and creating visualizations is time consuming and not the top priority for busy care providers. In addition, while some providers may prefer the freedom of creating their own visualizations, many would be overwhelmed by it.

**Solutions Comparison**

Making clinical decisions for HCC patients is a complex, difficult task. Table 1 summarizes the possible strategies for tackling this problem, along with the strengths and weaknesses for each approach. The current method of MDT discussions of patient data and BCLC staging guidelines is not prescriptive for all patients, nor are there opportunities for determining better guidelines, because data is unavailable. The MDBD tool currently in development at OHSU will collect HCC-specific data on a national scale, but it is still unclear how clinicians will best interact with this data for treatment decisions. Using existing business intelligence software such as Tableau or QlikView with the MDBD data will be too overwhelming and time-consuming for most care providers and its portability to other institutions might be limited. Therefore, clinical visualizations and visual analytics tools developed by informaticians and data analytics experts and informed by clinicians will provide the best opportunities for data interaction with the MDBD data.
Table 1. Comparison of approaches for HCC treatment decision making (Hribar et al. solution highlighted)

<table>
<thead>
<tr>
<th>Possible Solution</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online data visualization repository and clinical data visualization and visual analytics tool integrated with the MDBD application</td>
<td>Collaborative, flexible, iterative (see Proposed Solution section)</td>
<td>Requires providers take the time to review and evaluate visualizations; value of insights dependent on MDBD participation</td>
</tr>
<tr>
<td>Current MDT decision-making</td>
<td>Structure is currently in place</td>
<td>BCLC staging guidelines not complete for all patients; current HCC data not structured or centralized for possible learning</td>
</tr>
<tr>
<td>Using large national cancer databases such as SEER or CancerLinQ to inform decisions</td>
<td>Large EHR data repository, possible learning from national data</td>
<td>Not all currently in use; broad coverage; difficult to determine if necessary HCC data will be available</td>
</tr>
<tr>
<td>Using MDBD data with business intelligence software to inform treatment decisions</td>
<td>Providers are free to explore the national data to gain insights</td>
<td>Time consuming and requires data analytics skills; visualizations may not be useful; value of insights dependent on participation in using MDBD</td>
</tr>
</tbody>
</table>

Implementation and Dissemination

Initial development of the MDBD tool is underway within the Oregon Clinical & Translational Research Institute (OCTRI). Five regional medical centers will participate in this pilot program and will use the web browser-based MDBD tool in their liver tumor board meetings. Our proposed clinical data visualization and visual analytics tool will be implemented as a feature of the MDBD tool in a future version of the software. In parallel, our online visualization development and evaluation repository and forum will be made accessible to users. Members of the participating MDTs, clinical informaticians, and data analysts will be encouraged to participate in visualization development projects led by our group. The visualization development projects will include a pilot testing of the methodology as well as team member use of the data visualization and visual analytics tools in live MDT meetings.

After this initial implementation, we will invite other national and international HCC MDTs, as well as other disease-oriented MDTs, to participate in the visualization repository. In order to achieve wider dissemination, our solution will be presented to the medical community through scientific publications, presentations at national conferences, and other outreach venues.

Evaluation

Our proposed solution will be evaluated in four different stages: (a) user feedback regarding the data visualization prototypes in the online repository; (b) formal evaluation of the visualization solution using usability methodologies with clinician-users; (c) evaluation of the effect of the visualizations and visual analytic tools in the clinical decision-making, including clinician perception and satisfaction; and (d) evaluation of patient outcomes (Supplemental Materials, Exhibit 6).

Methods used for formal evaluation of the visualizations will include creating user-centered design principles, testing the visualizations in context, and evaluating information search strategy performance with real users⁶. Evaluation methods to assess the user’s analysis and reasoning capabilities⁷, as well as insight-based methodologies⁸ will evaluate usefulness of the visualizations. It will be important to measure the effects of a visualization on decision-making; we propose using an MDT teamwork measurement scale before and after using the visualization to measure this effect⁹. Finally, the evaluation of different patient outcomes (survival, remission, quality of life, time-to-therapy, etc.) will be performed with pre/post-intervention study methodology.
Conclusion

While demand is rapidly growing for visualization tools and techniques able to handle the complexity of clinical and biomedical “Big Data,” novel approaches to making these data easily accessible to and actionable by providers are in short supply. We believe applications like the HCC Data Visualization Repository and Clinical Tool have the potential to enable clinical insights at institutional and national levels, improve patient care, and support the development of a rapid-learning healthcare system.

References

Accelerating Biomedical Informatics Research with Interactive Multidimensional Data Fusion Platforms

Satyajeet Raje, MS\textsuperscript{1,2}
\textsuperscript{1}Dept. of Computer Science and Engineering, \textsuperscript{2}Dept. of Biomedical Informatics, The Ohio State University, Columbus, OH

Abstract

A prototype data fusion platform is presented that enables users to perform automated interpretation and ad-hoc integration of multiple datasets in real-time. It is designed to cut down data pre-processing time significantly. The prototype platform design illustrates interactive features that are highly visual and intuitive and do not require programming expertise. The demo is available at http://52.25.231.52:8080/dataFusionPlatform/.

Introduction

Data-driven research and analytics is at the core of clinical and translational informatics\textsuperscript{1}. A growing number of datasets are becoming publicly available in the current biomedical and healthcare research landscape. This, abundance of data has made large-scale data-driven biomedical research a real possibility. However, due to the lack of researcher-friendly tools for handling multi-dimensional datasets, data availability has not yet been effectively translated to its meaningful use.

Problem Description: Closer inspection of the current data-driven research process reveals that researchers spend more time in pre-processing and curating of data rather than the actual analysis itself. The amount of time and resources required to complete this pre-processing often limits the usefulness of the remaining research process. Moreover, the completely manual process renders handling of large and complex datasets or interpretation of multiple datasets in parallel nearly impossible for researchers. This has severely 'choked' the pace of research.

Quantitative data analysis involves several unavoidable pre-processing activities such as data exploration, collection, cleaning, curation, interpretation, etc. across diverse multi-dimensional datasets\textsuperscript{2}. These can be broadly classified into three steps of data discovery, interpretation and integration. The activities involved are as follows:

• **Dataset discovery** involves locating and selecting the required datasets and perform initial validity of the data contained within. This validation is based on its relevance to previous research and on provenance and quality information of the data.

• **Data interpretation** involves understanding what the data looks like at a highly granular (column) level and how it may be integrated with the other datasets that the researcher might be interested in. After finding a suitable dataset, researchers have to manually relate the columns from the different public and private datasets.

• **Data integration** involves combining selected datasets and tables based on the researchers’ interpretation. Currently, researchers perform this manually every time a new dataset is to be integrated. The task is finding possible “matches” across datasets and then performing the “join”. In database terminology a “join” can be defined as a point (usually a column) where data is integrated whereas a “match” is a possible join\textsuperscript{3}.

Informal interviews of biomedical informatics researchers revealed that most of these tasks are performed manually and thus are very time-consuming\textsuperscript{4}. The researchers also identified that though there are several data analysis tools readily accessible, there are very few automated tools to aid researchers in the pre-processing tasks. The difficulty and effort is amplified due to multiple iterations often needed in research.

Proposed Solution: Data Fusion Platform

For this submission, a proof-of-concept implementation of a data fusion platform is presented, designed to provide automated, end-to-end, real-time and ad hoc functionality across the data pre-processing tasks\textsuperscript{5}. The hypothesis is that such an automated “plug-and-play” solution would drastically reduce the pre-processing time, which, in turn would greatly accelerate the pace of biomedical and healthcare research.

The platform leverages reusable, semantically coherent annotations across datasets using standard reference ontologies in order to provide automated features for pre-processing needs of the researchers. This allows individual researchers, working independently to effectively crowd source dataset collections. Through the proposed platform, the researchers should be able to easily navigate through the dataset collection, interpret and integrate the relevant
datasets and obtain data in a format that is ready to be processed by various analysis tools (such as SPSS, Stata and R). The platform provides an intuitive and researcher friendly interface, without necessity of programming.

**Review of Comparable Existing Tools**

Several tools supporting data analytics activities in business environments are available. However, we argue that these are not suitable for academic environments. Let us consider why in detail. First, researchers working independently and do not share an integrated data store, as is the norm in large-scale business data analytics. So tools designed around centralized data structure and schema would not work in a research environment, which would benefit from a more distributed, crowd-sourced approach. Rather than pushing datasets into a unified predetermined schema, the researchers need a tool that performs quick, automated and interactive mapping and matching for ad-hoc interpretation and integration of datasets as needed.

Second, researchers often lack the resources of dedicated technological expertise that is available within enterprise data warehousing contexts. Thus, they might not have dedicated programmers, often required to work with commercial tools. Moreover, such tools are generally expensive and, thus, not accessible to individual academic researchers with limited research budgets.

From this understanding of the research process, it is clear that lightweight tools specifically designed for researchers that work with crowd sourced datasets are critical in order to accelerate the pace of data-driven research. Three types of tools are predominantly accessible to and used in data preprocessing by researchers, as identified by the interviews. These are data discovery tools, scientific workflow management systems and data integration platforms. In this section, a summary of the survey of the existing tools available for data preprocessing performed is presented.

*Data Discovery Tools:* These are often dataset collections or “knowledge stores” with search capability to query for relevant datasets. They tend to be domain specific, for example, the NCBI Entrez collection of biomedical datasets. Collections such as data.gov and census.gov for publicly available government datasets are especially interesting as they contain environmental and socio-economic datasets, crucial for cross-domain research. Thus, even though these tools are effective for discovery of topical datasets, the tasks of interpretation and integration are still manual.

*Workflow Management Tools:* Workflow-based data platforms, such as Taverna and Galaxy Workflow allow researchers to create and share common tasks that need to be performed on data in the form of series of repeatable steps in a single workflow. For example, the process of converting raw genetic data to a usable tabular format is a standard workflow. Such tools are useful to record common preprocessing steps such as transformations and cleaning on popularly used datasets and formats that are very domain specific. Thus, interpretation and integration of new datasets is still problematic.

*Data Integration Platforms:* As mentioned earlier, the commercially available data integration platforms are designed for large-scale shared business needs and not for the individual researcher. Traditionally, data integration follows the Extract-Transform-Load (ETL) process relying on schema mapping and matching. Even though automated matching techniques have been studied for long, they are still time-consuming and show moderate accuracies. Thus, they are not suitable for interactive real-time features and lack the ad-hoc integration functionality. So even though large-scale data integration platforms, such as Informatica and IBM’s Infosphere may be ideal for data analytics in a business environment, such an approach is not suitable for researchers’ needs.

**Design Methodology**

This section discusses the methodology followed while designing the data fusion platform in detail. An enterprise modeling approach was adopted. The steps involved are summarized below:

1. **Identification of purpose and scope:** Researchers working in the biomedical informatics and biostatistics domain were interviewed to understand the typical preprocessing workflow in quantitative data analysis. Much of this is discussed in the previous section. This allowed us to identify the specific problem that the data fusion platform should solve and define the scope.

2. **Identification of requirements:** They were then asked to identify functional as well as non-functional expectations from a hypothetical fully automated data fusion platform that aids in the preprocessing tasks. A grounded theory approach was applied to narrow down the requirements around the tasks of data discovery, integration and interpretation.

3. **Use case specification:** From the requirements identified above, use cases around the various data preprocessing tasks could be derived. These serve as the functional requirements of the platform. The data fusion platform implementation is expected to functionally address all of these use cases.
4. **Implementation**: The scope and use cases provide specifications for the prototype implementation. An agile methodology for rapid software development was adopted, which can accommodate changing requirements.

5. **Evaluation**: The platform is to be evaluated for both its functionality and usability. Functional evaluation involves task-based assessment of the system. The usability evaluation is designed using human computer interaction principles.

**Scope of the Data Fusion Platform**

Based on the responses of researchers, the scope of the platform can be defined as:

- Support bottom-up build up of crowd-sourced datasets, interconnected using semantic annotations that connect individual dataset schema to reference ontologies (such as MeSH, SNOMED, UMLS, dbpedia, etc.). This would enable moving away from predetermined centralized database schemas and promote efficient data sharing and reuse.
- The coverage of classes, entities, relationships, etc. by the reference ontology used for annotations is beyond the scope of the platform implementation.
- Provide a lightweight solution to handle research needs by replacing the overhead of schema matching and mapping and support automated interpretation and ad-hoc integration tasks for multiple / changing hypothesis.
- Aid automated and real time execution data preprocessing tasks with limited requirement of technical resources or programming expertise. The platform should be capable of accelerating data-driven research with an intuitive and interactive researcher friendly (and not programmer-oriented) interface.

**Functional Requirements Specification using Use Cases**

The interviewed researchers were asked to identify the expectations and specific functionality from a hypothetical data fusion platform for leveraging crowd-sourced datasets. These requirements were then narrowed down into specific motivating scenarios around the tasks of data discovery, integration and interpretation, illustrated below. These use cases serve as the functional design requirements for the data fusion platform.

**Discover relevant datasets and conduct initial assessment of suitability**

**Use Case 1**: "Find available datasets that can be integrated with given dataset"

Search for relational datasets of topical interest using existing data discovery tools like data.gov, census.gov, geodata, NCBI Entrez datasets, etc.

**Use Case 2**: "Get quality information about the dataset"

Assess provenance data (publishers, version, etc.) as well as quality information (e.g. number of data entries, analysis of missing values, etc.) of the datasets.

**Real-time Interpretation**

**Use Case 3**: "Find all possible matches between given datasets"

Automatically find "matches" between multiple local and public datasets using explicit table / column level annotations.

**Use Case 4**: "Suggest intermediate datasets for integration"

Automatically find intermediate datasets required to integrate given two datasets with no direct "match".

**Perform ad-hoc integration of data**

**Use Case 5**: "Suggest the best possible join for me to integrate given datasets"

Compare multiple datasets “matches” for relevance and integration possibilities in a real-time environment using heuristics of data quality / retention / loss. For example, datasets A and B can be joined on column X or can be joined using P, an intermediate dataset with a different set of columns. The number of data rows (observations) retained after the join can be a simple heuristic measure to compare which “match” is suitable for integration.

**Use Case 6**: Export the “joins” and data in format ready for analysis

Perform ad-hoc, real-time “joins” across identified datasets and deliver them in relational (tabular) format for use in various analysis tools like SPSS, Stata, etc.

**Non-functional Requirements**

The interviewed researchers also identified the following non-functional requirements for the data fusion platform:

- **Visual query interface**: An intuitive visual / graphical user environment that cuts across datasets, is not programmer oriented, and capable of delivering above functionalities.
- **Semi-automated annotation**: The platform should provide some degree of automation in the annotation process of a particular dataset. This would reduce the cost of annotation and improve heir precision.
- **Low query latency**: The query time for the automated interpretation tasks should be reasonably low to allow smooth operation in real-time.
• **Scalable Architecture:** The platform should be able to support a large number of datasets without losing the above requirements of intuitive interface and low latency.

**Implementation Details**

We have successfully implemented a prototype platform that fits the defined scope and addresses most of the stated requirements. It can perform automated interpretation across datasets and can identify “joins” for data integration using user-defined semantic annotations. The key design elements that help address the requirements are:

• **Data Fusion Ontology (DFO):** The internal storage ontology is key to the implementation of the platform. It defines the structure of the data and annotations stored within the platform. It is represented in an ontological format (OWL) that allows automated reasoning and extreme portability of the annotations.

• **Persistent data store:** A persistent data store is key to scalability of the platform. The annotations are stored in an ontological format (OWL / N-triples) using the DFO specifications. They can then be easily translated to any required database, for instance the Neo4j format for this implementation.

• **Graph-based exploration of datasets:** The interface allows an intuitive graph-based exploration of the datasets. The automated interpretation and ad-hoc integration can also be performed within the same interface.

The components of the platform are shown in figure 1. The dataset is input using the annotation interface, which interacts with the annotation engine. The user can annotate the tables and columns of the dataset using classes and properties from the reference ontologies. Note that in the above platform design the cost of annotation of a particular dataset is incurred only once. The annotation engine then verifies these annotations for completeness and validates it against the reference ontologies and the internal reference ontology, called the data fusion ontology (DFO). They are then translated into a Neo4j schema and loaded into the Neo4j graph database. These can now be reused by the query interface to perform the various preprocessing tasks as discussed in the use cases.

![Figure 1. Data Fusion Platform components](image)

For this design challenge, we are showcasing only the query interface with the assumption is that the datasets are already annotated and loaded into the Neo4j database. This is done in order to save the evaluators the time and effort of annotating new datasets. The demo interface is implemented as a web application using javascript and the D3 visualization library on the frontend and a Neo4j graph database at the backend. The database is queried using the cypher query language based on the user interactions at the front end. The interface allows users to browse the annotated datasets and perform real-time interpretation and ad-hoc integration. Currently, the platform can record and export the “joins” made by the users in the interface. This record can, in turn, be used to perform the actual integration. Please refer to the supplemental material (s1) for detailed walkthrough and screenshots of the demo design implementation.

**Large-scale implementation and Future development**

*Automated annotation process:* The annotation process is a major challenge in the large-scale implementation. The current implemented annotation process is completely manual. As part of future work, more automated techniques to aid the annotation process as shown in figure 1. These techniques would apply NLP methods as well as previous user annotations for automated classification of columns based on the reference ontology. Such a semi-automated annotation process should improve precision and drastically reduce the overhead of manual annotation.
Scaling the platform: There are two ways in which the platform is expected to scale. The most important consideration is how the platform will scale as more datasets are added expanding the size of the dataset collection. The current technology stack (viz. Neo4j, Cypher and D3) should be sufficient to handle this scaling. Another important consideration is allowing multiple users. Currently, the platform treats each session independently and all the changes and modifications are limited to the client-side only. Eventually, there would be need of storing user information and create profiles. This would provide better security through access control for the platform. It would enable advanced functionality such as user-specific collections, sharing, collaboration, etc.

End-to-end solution: The platform is expected to evolve as an end-to-end solution for data preprocessing. This would need inclusion of data discovery through search functionality. It would also require connecting to the actual data. We hope to expand on parallel research at OSU, such as ResearchIQ\textsuperscript{10}, in order to achieve this.

Evaluation Plan
As mentioned earlier, the data fusion platform will be evaluated for functionality as well as usability. We propose three different types of evaluation methods:

- **Task-based evaluation:** A task-based evaluation method involves identifying specific tasks to be performed by the system. In this case, the tasks will be designed around the use cases specified above. The platform will first be evaluated for the functionality by checking if it is actually capable of performing the tasks. A successful platform should be able to perform all the above use cases to be functionally complete.

- **Cognitive walkthrough:** The cognitive walkthrough can be considered as an extension of the task-based evaluation. An optimal workflow to perform each task is preredcorded. In the next stage of evaluation, a set of test users will be asked to perform the same tasks. The eyeball tracking and think-aloud protocols will be used as users are performing the tasks. An evaluator would also note various aspects of the process, such as whether the user could actually finish the task, the deviation from the optimal flow, time required to complete etc. Both the user actions and the responses will be recorded and analyzed to evaluate the user interface.

- **Heuristic evaluation:** The heuristic evaluation will be performed in two parts. First, the interface is to be evaluated by UI experts based on the HCI principles\textsuperscript{11}. Next, we would design a questionnaire based on these principles and to survey the naïve users of the system. The platform will be evaluated as a combination of the expert comments and the user responses.

Conclusion
The data fusion platform design presented here provides researcher with an intuitive interface to perform several preprocessing tasks. The implementation allows users to automatically interpret to find “matches” across datasets as well as perform ad-hoc “joins” to integrate datasets in a real-time environment. The platform is expected to evolve into an end-to-end solution for all data preprocessing tasks.

Acknowledgements
I would like to thank Justin Dano, Tyler Kuhn, Marques Mayoras and Bradley Myers who were instrumental in development of the prototype system. I would also like to extend a special thank you to my advisors Dr. Jayashree Ramanathan and Dr. Philip Payne. The demo is available at \url{http://52.25.231.52:8080/dataFusionPlatform/}

References
8. US Census Bureau, American Fact Finder, factfinder.census.gov
Three-dimensional Content-Based Cardiac Image Retrieval using global and local descriptors

Leila C. C. Bergamasco¹, MSc, Fátima L. S. Nunes, PhD²

¹Laboratory of Computer Applications for HealthCare, Department of Electrical Engineering - Polytechnic School, University of São Paulo (PPGEE-EP-USP)

²Graduate Program in Information Systems – School of Arts, Sciences and Humanities
University of São Paulo (EACH-USP)

E-mail: {leila.cristina, fatima.nunes}@usp.br

Abstract

The increase in volume of medical images generated and stored has created difficulties in accurate image retrieval. An alternative is to generate three-dimensional (3D) models from such medical images and use them in the search. Some of the main cardiac illnesses, such as Congestive Heart Failure (CHF), have deformation in the heart’s shape as one of the main symptoms, which can be identified faster in a 3D object than in slices. This article presents techniques developed to retrieve 3D cardiac models using global and local descriptors within a content-based image retrieval system. These techniques were applied in pre-classified 3D models with and without the CHF disease and they were evaluated by using Precision vs. Recall metric. We observed that local descriptors achieved better results than a global descriptor, reaching 85% of accuracy. The results confirmed the potential of using 3D models retrieval in the medical context to aid in the diagnosis.

1. Introduction

Technology has entered in all sectors of society over the last years, changing the way people work and interact. Medicine, in particular, is one of the areas that has most benefited from the technological advent and often leads the way in the use of existing advances.

Computer-Aided Diagnosis (CAD) schemes provide assistance on diagnosis, using patient data and medical image data. These systems highlight suspicious areas in a medical image, in order to provide detailed data from an anomaly to the health professional¹. Recently, Content-Based Image Retrieval (CBIR) has emerged as an important technology to aid diagnosis, since it allows retrieving from a database those images that are most similar to an image provided as an input.

Parallel to the technological advancement cited, the area of graphics processing, including image processing, computer graphics, virtual reality and related fields has also evolved over the years. One of the causes is because the hardware utilized by these areas has become more efficient and cheaper.

In this scenario, techniques have been developed for the reconstruction of three-dimensional (3D) models. One of the main reasons for the increasing use of 3D models in health is that they provide additional information when compared with two-dimensional (2D) images. Besides the information about depth and volume, it is possible to condense color, contrast and resolution characteristics from 2D slices³.

Because of these characteristics, many complex medical exams use 3D models in the diagnosis. Magnetic Resonance Imaging (MRI) and Computerized Tomography (CT) are examples of medical image modalities that allow experts to identify abnormalities such as aneurysms, coronary artery diseases and tumors in several internal organs of the body without the need for invasive methods³. However, each exam from these modalities can generate hundreds of 2D slices and a high volume of data to be examined by physicians in order to perform their diagnoses.

Cardiology can be greatly benefited by CBIR systems using these images. Through MRI and CT exams, computational aid can be provided for the analysis of different diseases such as cardiac ischemia, heart attacks and cardiac insufficiency. A possible solution to overcome the large data volume problem to be analyzed is to generate a 3D model from the 2D slices and use this model to aid the diagnosis. In the Information Retrieval area, the same approach can be used: the system uses the 3D model for extracting the features instead of hundreds of 2D slices. This solution can be faster and more accurate than the approach using 2D image², ³, ⁵. However, this approach also
has limitations, since the volume or the surface needs to be available. Obtaining this kind of object is not a simple process, since the slices need to be segmented and reconstructed.

In the health area the use of 3D models can be particularly useful for detecting shape alterations in the reconstructed structure. The shape deformation is an important symptom of many diseases, such as Congestive Heart Failure (CHF) which is characterized by the inability of the heart to pump blood at an adequate rate for the metabolic requirements of the body.

CHF can be discussed from two perspectives: Left-sided Heart Failure and Right-sided Heart Failure. CHF can affect only one side of the heart, and as it is a closed circuit, it is common that this insufficiency on one side makes the other side work harder, resulting in excessive strain, hence producing global CHF. This excessive strain produced in the ventricles can deform this structure in the mid- and long-term. Weight gain or thickness of the ventricle indicates hypertrophy, while increase of the chamber size indicates dilation.

In this context, a CAD system using CBIR concepts to assist the specialists in finding similar cases based on the 3D model of the structure analyzed can be quite useful both for the diagnosis and for medical education. Despite the large amount of retrieval systems for content-based 2D medical images, the research on 3D models is still incipient in the literature. It is therefore important to develop more specific techniques that take into account the particularities of medical models.

This article contributes to fill this gap by presenting the definition, implementation and validation of one global and three local shape feature descriptors, applied in reconstructed heart left ventricle models. These models were previously classified by an expert as cases with anomaly (CHF disease) and without anomaly. The aim of this research report is verify how different descriptors techniques behave on this specific scenario and which strategy is more precise.

This article is organized as follows: Section 2 presents some works related to the work theme; Section 3 describes the methodology used in the tests and describes the descriptors used; Section 4 discusses the results and Section 5 presents the final conclusion.

2. Background

Research on 3D CBIR is relatively recent, with the main articles of the area published over the last ten years. The components of a CBIR system, both in 2D and 3D domains, are descriptors, similarity function and indexation structures. Descriptors are algorithms that extract some feature from the images. A set of features forms a features vector. Similarity function is an algorithm (usually a distance function) used to measure how similar two images are. There are different metrics for calculating distances, such as Euclidean and Manhattan Distance. Indexation structures are data structures developed to make faster the retrieval.

In 2D CBIR the descriptors can be classified as global or local, and they are usually divided into three categories: shape, color and texture. Shape descriptors are predominantly developed in 3D domain, and there are numerous different subcategories that vary from author to author. Yubin et al. propose categorizing shape descriptors based on geometry, statistics or projection. The geometry category considers as basic features of the 3D model, such as vertices and volume. Descriptors that use quantitative analysis tools such as histograms to define the features and create descriptors are in the statistics category, and descriptors based on projection are those that extract 2D images from the 3D models and also analyze them by using 2D descriptors. We found only a study that performs the retrieval of 3D models based on their colors, which presented a limitation of retrieving 3D models with completely different shapes as being alike.

In relation to the difference between global and local descriptors, the former analyzes the 3D model as a whole, and the feature vector includes information from all regions of the 3D structure. Local descriptors analyze parts of the models individually. According to Qin et al., global descriptors are easier to implement and have shown robust results taking into account that the 3D models analyzed are simple models. However, for more complex and detailed models, using a local descriptor approach can be preferable. Although it can require more processing, retrieval is more accurate and makes the retrieval system more flexible, since it can select regions of interest to be compared.

The similarity functions applied to features vectors of the 3D CBIR systems analyzed are the same used by 2D CBIR systems, with a prevailing use of Euclidean and Manhattan distances. Since the features vectors are still a similar data structure, the type of similarity function used remains the same.
3D models of 3D CBIR systems cited in the literature are generally representations of domestic objects such as chairs, vases, bottles and cars. Benchmarks for 3D CBIR, as for instance the Princeton and McGill Benchmarks, also used generic models to evaluate their descriptor\(^5\). With this type of benchmarks global descriptors have presented good results, since these models have very distinct shape differences.

In the medical context, few works have focused on 3D CBIR. Glatard et al.\(^8\), for example, obtained a 3D volume of the myocardium, and use this object to identify the cardiac cycle phase (systole or diastole) a given query image was in and find slices similar to the query image. A projection-based descriptor that uses 2D images extracted from a 3D object was used.

In Wu et al.\(^9\) different volume descriptors were combined to analyze brain models from PET (Positron Emission Tomography) exams. According to the categorization of Yubin et al.\(^3\) these descriptors can be considered as geometry-based descriptors, since they analyze primitive information of the 3D model, such as the volume and number of voxels. Finally, Aman, Yao and Summers\(^10\) used the SIFT descriptor and Bag of Words to retrieve Colonography CT Scans. The authors applied the Normalized Discount Gain metric to evaluate their results\(^10\).

The development of CBIR systems based on 3D models, in order to aid the diagnosis, requires knowing the anatomical structure of the organ in the images and the anomaly under study. Often, the global shape descriptors have limitations, requiring local descriptors, as described in this present work.

### 3. Methodology

To achieve the proposed objectives one global and three local descriptors were defined and implemented. To calculate the similarity among the models the Euclidean Distance was used, as shown in Equation 1, where \(x\) and \(y\) are two feature vectors with \(n\) size and \(i\) indicates the \(i\)-th position at each feature vector. In the next sections each of the descriptors implemented will be detailed.

\[
\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)
\]

#### 3.1. Descriptors

**3.1.1. Distance Histogram Descriptor (DHD)**

The Distance Histogram Descriptor (DHD)\(^12\) considers the surface and the geometry of the 3D model analyzed. The algorithm computes the distance between the centroid of the model and its surface considering random points. Next, the distance is divided into ranges that form the bins of a Distance Histogram.

Figure 1 shows the DHD adaptation for the context of this work. From a reconstructed model provided as query, the distance between the centroid of this model and the random points on the surface is calculated. This distance is normalized, and the frequency of occurrence of each distance value in the model is computed to compose a histogram. This histogram is transformed into a feature vector and compared with other vectors stored in the database. To compute the similarity between two vectors the Euclidean distance is applied and the models are classified according to the value obtained. The lower the value, the more similar the model stored in the database and in relation to the query provided\(^13\).

**3.1.2. Local Distance Histogram Descriptor (LDHD)**

As mentioned earlier, CHF may cause ventricle deformations, especially in the lower regions of the structure. This descriptor takes in account this information and analyzes specific parts of the model based on their octants.

First, the 3D model is divided into eight octants and then the Distance Histogram of each part is calculated. For each Distance Histogram created, its respective area is computed, as given by Equation 2, where \(f_i\) is the \(i\)-th value of frequency, \(d_i\) is its respective value of distance, and \(n\) is the number of histogram bins. In other words, each of the eight positions of the distance vector stores the area of the \(i\)-th octant (Figure 2).

\[
\text{area} = \sum_{i=1}^{n} f_i d_i \quad (2)
\]

Figure 2 shows an example of each step of the LDHD. From a model reconstructed and used as query, the Distance Histogram is calculated for each of the octants, considering the centroid of the model as the point of origin for all of them. The area of each histogram is calculated and stored in a vector. This vector is compared with other vectors in the database using a similarity function. If the result is less than a threshold value, the result for this comparison is
zero, otherwise it is 1. The sum of the end positions of this vector indicates the degree of similarity between the two models.

![Figure 1](image1.png)

**Figure 1.** Steps performed by Distance Histogram Descriptor.

The problem of using common similarity functions directly into features vectors in this case is the loss of information about the octant reference. Most of the similarity functions cited in the literature do not consider the positioning of each feature in the vector; all features have the same weight and if one of them has a bad result all the 

![Figure 2](image2.png)

**Figure 2.** Steps of the Local Distance Histogram Descriptor.

The problem of using common similarity functions directly into features vectors in this case is the loss of information about the octant reference. Most of the similarity functions cited in the literature do not consider the positioning of each feature in the vector; all features have the same weight and if one of them has a bad result all the
feature vector can be compromised. Thus, the main advantage of this descriptor regards allowing the end result of the descriptor to reflect the local differences of the model analyzed. This type of approach in the 3D scenario is new, because it allows applying a global solution at specific points in the model.

3.1.1. 3D Hough Transform Descriptor (3DHTD) - Frequency

The Hough Transform is an effective technique for detection of curves and objects in 2D and 3D domain from a set of points. An important characteristic of this technique is how it discretizes the spatial information of the image in order to identify which points belong to the same set of interest\(^4\).

R-Table is a reference table in which the indexed curve is represented by the angle of the gradient and which is widely used in 2D Hough transform\(^4\). In the 3D domain, R-Table can be used to organize information about the normal vector, which indicates the positioning of a given surface, in addition to the distance of this surface related to the origin. Compared to other descriptors that only take into consideration the distance, there is a major gain for spatial problems that require detecting shape alterations at specific locations.

In the 3D Hough Transform Descriptor (3DHTD) descriptor, the discretization of spatial information occurs by using a Cubic Matrix, a 3D extension of R-Table, where the values of the rows and columns are measured according to a predetermined degree of resolution based on spherical coordinates \((\theta, \varphi)\). This degree of resolution indicates the range value on cubic matrix that the information will be grouped. Thus, lower resolution values indicate a larger area to be analyzed. The normalized Euclidean distance between the centroid and surface is computed to measure the depth of the Cubic Matrix. Finally, the content of each cell is defined by the frequency that this triple \((\theta, \varphi \text{ and distance})\) occurs in the model.

The set of triples of a model, along with its respective frequency of occurrence \((\text{freq})\), is stored in the database as a position in the feature vector of the model, as shown in Equation 3, where \(a_i\) is the \(i\)-th Cubic Matrix set \(\{\theta, \varphi, \text{distance, freq}\}\).

\[
\text{FeatureVector} = \{a_1, a_2, ..., a_n\} \quad (3)
\]

The similarity between two models is obtained by comparing the frequency of each cell of the Cubic Matrix. Figure 3 shows how this descriptor works: from a 3D model given as query the information is extracted using 3DHTD – Frequency, and the Cubic Matrix generated (mRquery) is compared by Euclidean Distance with another Cubic Matrix stored on Database (mRDatabase).

![Figure 3](image_url)
3.1.3. 3D Hough Transform Descriptor (3DHTD) – Standard deviation

This descriptor is a second approach of 3DHTD presented in the previous section, which analyzes the distance variations of the surface up to the origin point of the model in the spatial intervals defined by a given degree of resolution.

To reduce the dimensionality of the Cubic Matrix in order to find the standard deviation values, the mean distance for each pair of angles \((\phi, \theta)\) is computed. Next, the standard deviation of a set of frequencies using Equation 4 is calculated, where \(f_i\) is the \(i\)-th frequency of the pair of angles \((\phi, \theta)\), \(\mu\) is the average of the values found (in this case it is the distance), \(x\) is the cell value and \(n\) is the total of occurrences within the range of each pair of angles \((\phi, \theta)\).

\[
DP = \sqrt{\frac{\sum (f_i(x-\mu)^2)}{n}} \quad (4)
\]

Afterwards, Euclidean distance is applied, which verifies the entire matrix by comparing the values of the standard deviation, as shown in Figure 4. Similar to 3DHTD – Frequency, the feature vector is created from a 3D model given as query and a matrix based on Standard Deviation (matrixSDQ) is generated. This matrix is compared with another matrices stored on database (matrixSDR) by using the Euclidean Distance.

![Diagram of 3D Hough Transform Descriptor using standard deviation](image)

**Figure 4.** Implementation of the 3D Hough Transform Descriptor using standard deviation.

3.1. Experiments

The descriptors presented in the previous sections were applied in models of the left ventricle reconstructed from MRI images. Each MRI exam was composed by 45 heart slices obtained during diastole. These slices have spatial resolution of 256x256 pixels and contrast resolution of 16 bits per pixel. Besides the Precision versus Recall curve, the response time of each of these descriptors was also analyzed.

To test the descriptors in medical images, 30 sets of MRI exams from Heart Institute (InCor) were used - 53% exhibited the problem of CHF and 47% showed no abnormality. 55% of patients were women and 45% were men.
The frames were segmented focusing on the region of interest in the left ventricle using Seg3D software\textsuperscript{15}. The ImageVis tool was used for the reconstruction\textsuperscript{16}. Figures 5 and 6 show an example of the slices and the reconstructed ventricle, respectively.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{image1.png}
\caption{Example of a slice and the ventricle region segmented (in red).}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{image2.png}
\caption{Examples of reconstructed ventricles.}
\end{figure}

We used the Precision versus Recall metric to evaluate the results of the CBIR system, where Precision indicates the proportion of relevant images retrieved and Recall indicates the proportion of all relevant images that are in the database and that were retrieved\textsuperscript{18}. We also used a Distance Matrix, where the cell color indicates how similar a model is to the model given as query. The matrix was divided into two clusters: one related to patients with CHF (patients 1 to 16), and another related to patients with no anomaly (patients 17 to 30). The color of the cells indicates the order in which they were retrieved: the darker cells indicate lower distances and, therefore, models were retrieved before the models related to the lighter cells.

4. Results and Discussion

Figures 7 and 8 show the Precision versus Recall curve for each descriptor. In both cases, with and without the CHF disease, the 3DHTD using frequency was the best performing descriptor, indicating that spatial information is important for retrieving objects that have specificities at different places of the models. This descriptor reached an average of 85\% of precision for lower values of recall, which means that as the number of models retrieved increased, the precision decreased, since we have fewer models available on database and the shape difference between them becomes more difficult to distinguish.

The 3DHTD descriptor using standard deviation had less satisfactory results. The main limitation of this latter method is that small deformations may not be taken into consideration due to the standard deviation characteristic which considers the average of the range of values to compute the deviations of the data set.

The LDHD, which divides the models into octants and creates an auxiliary vector to compare local distances, presented good performance, about 75\% of precision, showing that it is feasible to use locally the global descriptor Distance Histograms. The Distance Histograms, in turn, showed a somewhat less satisfactory performance, about 60\% of precision, mainly due to the specificities of the models, in which local information is extremely important for composing the diagnosis.
Figure 7: Precision vs. Recall curve comparing all descriptors implemented in the models with CHF.

Figure 8: Precision versus. Recall curve comparing all the descriptors implemented in the models with no anomalies.

The errors observed for the first five models retrieved for each query can be found in the Distance Matrices (Figure 9). The 3DHTD - Frequency and the LDHD descriptors had a better performance that can be realized by comparing the generated clusters. DHD and 3DHTD – Standard Deviation descriptors had many outliers, while 3DHTD - Frequency and the LDHD descriptors generated more uniform clusters, separating into two groups the 3D models with and without the CHF disease.
These results show that the descriptors that take into consideration the spatial location of deformations achieved better performance than the descriptor with a global approach. The local descriptor 3DHTD – Frequency, for example, had a better performance — above 20% to 30% more precise — compared to Distance Histogram that uses a global approach. 3DHTD uses spherical coordinates and the distance of the face to the centroid to extract features from the 3D object. Consequently, it considers spatial location as well as the deformation degree as elements of comparison. Additionally, the 3DHTD – Frequency compare each model according to the deformation pattern. In other words, if two models have the same deformation on the same location but their frequencies are different, this means that they are different because the intensity of the deformation are highlighted by the frequency. The performance was consistent with the theory in which the descriptors were based on. Given that the main problem was to identify local deformations, the descriptors that provided this type of information and that cross-referenced with other data such as the degree of deformation (indicated by the distance from the surface to the center of the model) were more likely to achieve better results.

In this work the Euclidean distance was used, however there is a large set of other functions that can be implemented. Thus, evaluating results using other distance functions and develop more descriptors within this context can also enrich the discussion about 3D CBIR, hence offer a contribution to the actual state of the art. The current studies that we found in the literature usually use the same strategy applied in the 2D domain, commonly distance functions — as Euclidean and Manhattan — that measure the distance between two feature vectors. In this paper, for example, we have a 3D feature vector – the Cubic Matrix, and complex data structure to store the information resultant from the descriptors which we created. This can be considered a new way to store and compare the huge volume of information provided by medical images.

5. Conclusion

The objective of this paper was to define, implement and validate four descriptors and compare their performance for the retrieval of three-dimensional medical models with deformations at specific locations, in order to aid the CHF disease diagnosis. In many medical imaging exams it is common that several slices of a structure have to be analyzed by the expert. The generation of 3D models reconstructed from these slices can be a way to decrease the amount of information to be analyzed by experts. In addition, the development of descriptors that can characterize adequately these models is an important contribution for the computer-aided diagnosis area. As mentioned in the Discussion section, local descriptors achieved better results: they were above 20% to 30% more precise than global descriptors, mainly because we were looking for retrieve 3D models with specific and small changes.

This study has some limitations. The first of them is the low number of cases to test the developed techniques, which became difficult to make a deeper analysis of the results considering, for example, gender and age. Due to this limitation, it was also difficult to identify more specific variations on each group. The second one was the use of only Euclidean distance as similarity function; other metrics and evaluation forms can be explored.

Thus, from this work we show evidences that it is feasible to apply the concepts of 3D CBIR for more specific contexts. In this paper CHF disease was studied, however the descriptors developed in this paper can be applied to solve other problems that include local deformations of 3D convex models. In addition, the application of these descriptors in non-convex models also can contribute to evaluate them in a new context. New topics for study about 3D CBIR system can also include similarity functions for 3D feature vectors and complex structures. For future
work we are working in increase our database with more 3D medical models. We also are planning to apply these descriptors as well as others new techniques to retrieve cases with different cardiac diseases with a high level of precision.

As mentioned, there are currently few studies related to CBIR in the 3D domain and, therefore, this research presents a contribution comparing performance of global and local descriptors facing the recovery of 3D medical models for aiding the diagnosis of CHF disease.

Acknowledgement
This research was supported by the State of São Paulo Research Foundation (FAPESP - Process #2010/15691-0 and 2011/15949-0), Heart Institute (InCor), Brazilian National Council of Scientific and Technological Development (CNPq - Process #559931/ 2010-7 and #401745/2013-9). The National Institute of Science and Technology Medicine Assisted by Scientific Computing (INCT-MACC) and we thank Prof. Dr. Carlos Eduardo Rochitte for your time and expertise.

References
Analyzing Self-Help Forums with Ontology-Based Text Mining: An Exploration in Kidney Space

Philipp Burckhardt, MSc, Rema Padman, PhD
Carnegie Mellon University, Pittsburgh, PA

Abstract

The Internet has emerged as a popular source for health-related information. More than eighty percent of American Internet users have searched for health topics online. Millions of patients use self-help online forums to exchange information and support. In parallel, the increasing prevalence of chronic diseases has become a financial burden for the healthcare system demanding new, cost-effective interventions. To provide such interventions, it is necessary to understand patients’ preferences of treatment options and to gain insights into their experiences as patients. We introduce a text-processing algorithm based on semantic ontologies to allow for finer-grained analyses of online forums compared to standard methods. We have applied our method in an analysis of two major Chronic Kidney Disease (CKD) forums. Our results suggest that the analysis of forums may provide valuable insights on daily issues patients face, their choice of different treatment options and interactions between patients, their relatives and clinicians.

1 Introduction

In recent years, the Internet has emerged as a major source of health-related information, with more than 80% of the US Internet-user population, 113 million adults, having used the Web this way. How health information on the Internet affects decision making has been studied thoroughly as well as the demographics of help seekers, yet another area has received much less scrutiny: The proliferation of online self-help forums for chronic diseases that facilitate a social support network for patients.

The reason for this lack of attention is two-fold: On the one hand, the text amassed in these forums is generally too large to be hand-coded and subjected to a qualitative, human-guided analysis. On the other hand, automated text mining with the goal of discovering knowledge is a very challenging task and an active area of research.

In natural language processing (NLP), there has been a lot of work done in areas such as sentiment analysis, document summarization and topic modeling. Most of the current approaches are unsupervised learning methods, which utilize elaborate statistical methods and require a large amount of training data in order to generate sound results.

Although we are living in the era of ‘Big Data’, the data sets of these specialized self-help forums are much smaller than those of popular social networks. Confronted with this challenge, we opted to create a method for analyzing textual documents that deals with the relative data scarcity. In addition, popular topic modeling algorithms such as Latent Dirichlet Allocation (LDA) are mostly used to cluster documents into broad categories. Yet, our goal is to uncover hidden facts that would otherwise remain unknown.

In our proposed method, we use semantic ontologies to annotate natural language documents and mould them into a machine-readable format, with the goal of encoding the ontological knowledge a human might have. Unlike most statistical methods which use word counts as their atomic units, our representation is based on inter-related semantic concepts, which form an intermediate ontological layer representing a ‘web of meaning’. This ‘web of meaning’, which consists of hierarchically organized conceptual nodes, is then analyzed by statistical means, the results of which can inform and ease content analysis.

Because of aforementioned challenges, not much research has been done in the analysis of self-help forum communication. Existing studies have investigated the impact on attitudes towards healthcare providers running such forums. Others have studied the benefits and challenges arising from the usage of online communities by patients.

Instead of engaging in the debate on the perceived benefits and disadvantages of such forums, we take the view that even erroneous conceptions and misunderstandings on the side of the patients provide valuable insights insofar as they manifest themselves in the minutiae of online discussion forums and can thus inform researchers on how chronic disease patients deal with their conditions. Taking this perspective, Liu et al. study how such patients use video blogs.
to share their stories, and Zhang et al. investigate the intent of people posting on health message boards. There is also a small but growing body of research on how to evaluate the benefits and harms of different treatments and drugs by utilizing comments on online medical forums.

The results of analyzing online forums can lead to a broadened perspective for the healthcare sector. Let us assume a doctor who treats a patient with a chronic disease. Not only has there been a decline in the time doctors can spend with their patients, but due to the nature of the doctor-patient relationship, which is confined to single points in time, the doctor only gets a snapshot and not the whole picture of the patients’ health. Therefore, it would be desirable for a clinician to learn how the patient feels, whether the treatment is successful and what other issues the patient might have. The analysis of forum communication of patients facing similar problems could extend the vision of the attending clinicians and increase their understanding of patients’ needs and preferences.

Discerning communication patterns on self-help networks could also give rise to insights on how patients complement the information received from their doctors. These insights could lead to suggestions to improve the guidelines for doctor-patient communication and to integrate self-help networks into treatment programs.

In 2011, $49.2 billion was spent in the United States on the treatment of End Stage Renal Disease (ESRD), the final stage of CKD. It is estimated that more than 11% of the US adult population have some degree of CKD, with recent projections suggesting that more than 50% of those aged 30 to 64 years will likely develop CKD. In light of these trends, seeking improvement in the care of CKD patients has become a national issue as well as an economic necessity, both in terms of developing new treatments as well as preventive measures. This challenge is not restricted to the US, with CKD being called a “global challenge” requiring concerted action “to avoid a major catastrophe.” Although the gained insights might be very valuable, we are not aware of any studies learning communication patterns and insights from freely available CKD self-help forums - there has been only a small-scale intervention study targeted at adolescents. In this study, which serves also as a use case for our methodology, we report on the results of an analysis of two major CKD self-help forums, DaVita and KidneySpace. The insights we have gained in this area render the pursuit of our methodology a promising endeavor.

2 Data

The basis of our analysis is formed by the communication acts from the following self-help forums about CKD: DaVita, a large forum on kidney disease and dialysis hosted by DaVita, one of the primary kidney care companies in the United States, and KidneySpace, a privately operated kidney support forum that was run by the Renal Support Network until it closed down on January 30, 2015.

Table 1: Example thread from KidneySpace forum in JSON format

```json
{
  "posts": [
    {
      "content": "With all the attention around transplant, UNOS has opened up a number for people to call. (…)",
      "title": "\# to call regarding concerns about your transplant center",
      "user": "OZfan",
      "date": "August 17, 2007, 03:17:18 PM PDT"
    },
    {
      "content": "Hopefully I won’t have to use this number.",
      "title": "Re: \# to call regarding concerns about your transplant center",
      "user": "Purple_Reign",
      "date": "August 23, 2007, 12:57:49 AM PDT"
    }
  ],
  "title": "\# to call regarding concerns about your transplant center",
  "url": "http://www.kidneyspace.com/index.php/topic,40.1"
}
```

For each thread, our data set contains the following features: Its title, URL address as well as a list of posts, which
includes both the content of the posts as well as information on the author and date the post was written. An example thread is displayed in Table 1.

In total, our data set of the Davita forum consists of 6,501 threads with 41,079 posts. The time of the posts ranges from 2004 to the beginning of 2014, for a total of almost ten years. For KidneySpace, we have collected 3,715 online discussions totaling 20,857 posts from August 2007 until the year 2013, with 581 unique users present in the data set. For Davita, 5,603 users are represented in the study. This number is striking insofar as the total number of registered users exceeds 180,000, implying that a very large number of registered users did not write a single post. This trend continues if we look at the distribution of the number of posts per user, which are displayed for both forums in Figure 1. It seems as if most of the users are mere bystanders who do not actively communicate on the discussion forums, but instead are passive readers. This practice of *lurking* is commonly observed and has been studied in the literature.\footnote{23}

<table>
<thead>
<tr>
<th># of Posts</th>
<th>User Count</th>
<th>Post Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Davita</td>
<td>KidneySpace</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\begin{figure}[h]
\centering
\includegraphics[width=0.4\textwidth]{davita.png}
\includegraphics[width=0.4\textwidth]{kidneyspace.png}
\caption{Post frequencies for users of both forums. The number of posts has been censored at a value of 25, with the last bar in both plots denoting users with more than 25 posts.}
\end{figure}

Comparing the two forums, it can be seen that KidneySpace, although a much smaller forum in total, seems to have a more active community: The average number of posts per active user is 2.588 for Davita and 6.196 for KidneySpace.

Before our data analysis, we pre-process the raw forum data and annotate it with ontological information, turning it into a structured, machine-readable and statistically analyzable corpus. The necessary background information and the proposed methodology is presented in the next section.

3 Methods

Traditionally, statistical models for text analysis use word counts as their atomic units. Such a representation, which is devoid of meaning, is ill-suited to the characteristics of languages, which often display *polysemy* and *synonymy* of words.

To illustrate the problem of synonymy, we may use a mundane example: Every native speaker of English knows that the expression “Give me a buck” is equivalent to “Give me a dollar”, which in return is equivalent to “Give me a clam”. Any word-based model will treat the three words buck, dollar and clam as different entities, and there would be no inherent connection between them. A human listener though recognizes that these three words are synonyms and share a common meaning. However, not only can multiple words refer to the same concept, but a word can have different meanings, too. This polysemy may be illustrated by the word ‘lemon’ which denotes a fruit as well as a defective automobile.

When we as human beings assign the proper meaning in a given context, we make use of our ontological knowledge. To train an algorithm to make such an assignment, there is a need for a digital ontology. For this project, we have evaluated three commonly used ontologies: ConceptNet, the UMLS Semantic Network and the WordNet taxonomy.\footnote{24, 25, 26} In order to use these ontologies, we wrote APIs for each of them and made them publicly available under an open-source license\footnote{UMLS: http://bit.ly/1BjRuiM, WordNet: http://bit.ly/1MLW421, ConceptNet: http://bit.ly/1C2tmDD}.

Although our proposed methodology is ontology-agnostic, we have settled on WordNet as it does not have some of the
Figure 2: Synsets of the noun ‘doctor’ and their hypernyms

disadvantages of the other two options. ConceptNet, developed at the MIT Media Lab, aggregates information from various sources, at the cost of structural inconsistencies. The Unified Medical Language System (UMLS) contains a vast dictionary of more than two million biomedical and health related concepts (called the Metathesaurus), but its general ontology is very flat, consisting of merely 133 inter-related entities, whereas WordNet establishes relationships between all of its 117,000 concepts, called synsets, an abbreviation for sets of synonyms. However, a potential issue with WordNet might be its missing specialized medical concepts. Therefore, we thought of enhancing it by UMLS. Since UMLS does not contain colloquial language needed in our context, we considered the Consumer Health Vocabulary\(^2\), which provides a mapping from colloquial terms to the relevant concepts in UMLS. But since we found that approximately 84\% of words could be assigned to a WordNet synset, the marginal benefit from incorporating UMLS information into WordNet seems limited.

The advantage of mapping words to their respective ontological concepts is that the relationships between these concepts enable us to attach background information to our data. We focus on the Is-A relationship between concepts. This results in a taxonomy that can be represented as a tree structure with very general concepts on the higher layers of the tree and very specific concepts as leafs. For two concepts sharing an Is-A relationship, we refer to the broader concept as the hypernym of the more specific concept. In Figure 2, the possible synsets for the noun ‘doctor’ are displayed as well as their hypernyms. Because of this tree of knowledge, we can deal with the linguistic problems of polysemy and synonymy.

**Pre-Processing:** We start by tokenizing each input document and then run a standard Part-Of-Speech (POS) tagger to infer the lexical class of each extracted token. We remove interpunctuation, convert characters to lower case and strip off extra whitespace occurring in the documents. We then exclude all stop words from the analysis. Our system uses a stopword list of 571 words, including conjunctions such as and, auxiliary verbs such as should and pronouns like they. After removal of stop-words, we use a custom-implementation of WordNet’s morphy algorithm for morphological processing in order to map inflected forms back to their base forms. One of the input parameters of the morphy function is the POS tag, which identifies words as a noun, adjective, verb or adverb. Because of this intermediate step, it is possible to associate each word with all candidate synsets from the ontology, i.e. all possible meanings the word could have. Words for which no base form could be inferred are removed from further analysis. The steps of the pre-processing stage are visualized in Figure 3.

**Word Sense Disambiguation:**

The goal of *word sense disambiguation* is to choose the appropriate sense of a word given its context. For each word, we wish to select the correct synset \( c \) from the candidate set \( C \) of possible synsets. For example, when the word

\[\text{See: } \text{http://www.consumerhealthvocab.org/}\]
Figure 3: Pre-Processing Steps

Eating too much protein can strain the kidneys.

Figure 3: Pre-Processing Steps

The figure shows a flowchart of the pre-processing steps for text analysis. It includes steps such as tokenization, part-of-speech tagging, stopword removal, and morphological analysis. The flowchart illustrates the transformation of text into structured data for further analysis.

‘lemon’ occurs in a text about fruits, we would like to choose the synset with the definition *yellow oval fruit with juicy acidic flesh*, whereas in a text about used cars it is highly probable that ‘lemon’ refers to the synset \{lemon, stinker\} denoting an artifact (especially an automobile) that is defective or unsatisfactory.

Similarly to the approach by Fodeh et al.,\textsuperscript{27} for any given word \(w_i\) and its context \(W(i)\) (in our case, the other words in the sentence of the \(i\)-th word), we choose the synset \(c_i\) as

\[
\arg \max_{c_i \in C_i} \sum_{w_j \in W(i)} \max_{c_2 \in C_j} \text{sim}(c_1, c_2),
\]

where \(C_i\) and \(C_j\) are the sets of possible synsets for words \(w_i\) and \(w_j\) and \(\text{sim}(c_1, c_2)\) is some similarity measure between the synsets \(c_1\) and \(c_2\). The idea behind this equation is familiar to anybody: That the narrative context gives each word its special meaning.

In the literature, many different similarity measures have been proposed,\textsuperscript{28} some depend on the taxonomy and deem two synsets to be more semantically similar the closer the two nodes are together in the hyponym tree,\textsuperscript{29} other so-called Information Content metrics take probabilistic information of synset occurrence into account.\textsuperscript{30} Jurafsky and Martin argue\textsuperscript{28} that of the similarity or distance based information measures, the Jiang-Conrath similarity measure\textsuperscript{31} performs most consistently. This measure takes both taxonomic and probabilistic information of synset occurrence into account. Inserting the definition of the Jiang-Conrath similarity measure into Formula (1), we estimate the synset for word \(w_i\) as

\[
\arg \max_{c_i \in C_i} \sum_{w_j \in W(i)} \max_{c_2 \in C_j} \left[-\text{IC}(c_1) - \text{IC}(c_2) + 2 \times \text{IC}(\text{les}(c_1, c_2))\right],
\]

where \(\text{IC}(c)\) denotes the self-information of concept \(c\), a quantity defined as \(\text{IC}(c) = -\log P(c)\), \(P(c)\) being the probability to observe an instance of concept \(c\). We choose the context \(W(i)\) as the sentence in which word \(w_i\) appears.

In the formula, \(\text{les}\) denotes the lowest common subsumer, the lowest node in the hyponym tree which is an ancestor for both \(c_1\) and \(c_2\). This formula reflects two forces: On the one hand, concepts are more likely to be chosen the more frequent they are. On the other hand, the information content of the \(\text{les}\) has the effect of favoring synsets that share a very specific \(\text{les}\): For example \{nurse\} and \{doctor\} have \{medical practitioner\} as \(\text{les}\). Its information content is higher than in the case where the \(\text{les}\) for both synsets would be just \{entity\}. Thus, in co-occurrence with ‘nurse’, we conceive the ‘doctor’ as a \{medical practitioner\} and not as a \{intellectual\}.

In order to compute \(\text{IC}(c)\), we estimate the probabilities \(P\) via their empirical frequencies. Obtaining synset counts of a document corpus is often prohibitively costly, as it would require to manually annotate each word with its correct synset. A common practice is not to insist on obtaining a sense-tagged corpus and instead calculate the empirical frequencies by summing up the number of occurrences of words belonging to synset \(c\).\textsuperscript{30} This method is undesirable for disambiguation, as its only effect will be that synsets with multiple words end up having a higher probability of being chosen. Instead, we make use of the Brown corpus, one of the few corpora for which the occurrence of synsets has been determined by annotating its documents by hand. However, the corpus size is only 500 documents, so the Brown corpus has the drawback that many synsets are assigned a count of zero. To increase disambiguation performance, a straightforward approach is to manually annotate a training set from the domain of interest, in our case
Figure 4: Simplified Illustration of the Merging Process

medical self-help forums on CKD. In order to make this process easier, we have developed a command-line tool where users can manually annotate words in training sentences with their correct synsets. This allowed us to increase the percentage of correctly disambiguated synsets to around 75%.

**Document Merging and Word Propagation:** Each document is now represented as a set of disambiguated synsets. Since we have augmented each synset with ontological knowledge in form of its hypernyms, it has now the form of a document tree. Next, we merge all document trees together into a single corpus tree (see Figure 4). In doing so, we take all words that have occurred in the corpus and attach them not only to their respective synset, but also to all its hypernyms. To make this clear, consider the example of a ‘nurse’ which belongs to the \{nurse\} synset. However, a \{nurse\} \textbf{Is-A} \{health professional, health care provider, caregiver\}, and the word propagation algorithm ensures that the word ‘nurse’ is also attached to the hypernym synset. This allows us to pose questions such as ‘Which words denoting health professionals appear in our corpus?’ and obtain answers that also include all words which belong to children of \{health professional\}.

**Thresholding:** Many synsets appear only once or twice in a document corpus and could be the result of random noise. We want to discard these synsets and only keep those that are reflective of the inherent characteristics of the document collection. As a default choice, we remove all synsets from the tree which appear in less than \(\sqrt{n}\) of the documents, where \(n\) denotes the total number of documents.

**Analysis:** The analysis of the generated synset tree proceeds in three different steps. First, we have a global view on the univariate frequencies of the concepts. From this most distant zoom level, we have a bird’s eye view that allows us to detect the most probable and conspicuous synsets. The ontological tree structure in its visual form allows us to easily explore the corpus, for example by starting from highly abstract concepts drilling down to more specific ones. To give an example, we might encounter the abstract synset \{emotion\}. Looking what kind of emotions appear, we would see that \{cravings\} appears surprisingly often, and that the word most often disambiguated to this synset is ‘appetite’. A likely hypothesis might be that articulated desires are almost exclusively linked to nutrition.

For the identified synsets (in above example, nutrition or cravings), we take a detailed view investigating their Pearson correlation with other synsets. This allows us to identify interaction fields between different synsets and sharpen our hypotheses. Because we are looking for real correlations and not chance occurrences, we have to carry out hypotheses tests. Here, we also have to account for multiple testing since for any given synset, all possible pairwise correlations with other synsets (for which there is no word-overlap) are calculated. We do this by controlling the false discovery rate (FDR), i.e. the expected percentage of correlations wrongly considered significant, at level 5%. Specifically, we use the procedure by Benjamini-Yekutieli as it is applicable even under arbitrary dependencies among tests. In our example, we find to our surprise that the most correlated synset with \{cravings\} is \{nausea,sickness\}, raising the question whether the patients are not actually complaining about a loss of appetite.

In a final step, we check the formulated hypotheses by tracking the proportion of posts in which they can be established.
To do this, we draw a simple random sample from the documents in question (a size of 100 is chosen as a default), in our example documents containing remarks about ‘appetite’. Selecting a random sample helps in eliminating potential biases by giving all documents an equal probability to be selected.

A human judge then goes through automatically generated summaries of the selected documents. This human-guided step is crucial as there are often multiple explanations for the observed frequencies and correlations, and only a look at the actual documents can reveal which of them apply.

4 Results

CKD patients eventually have to face a decision regarding which treatment to undergo. The following options exist: Hemodialysis (HD), peritoneal dialysis (PD) and kidney transplantation. The different treatment options all have their own risks and benefits, so patients might not be sure which suits them best. In addition to that, patients might have to change treatment because a transplant is rejected and has to be replaced by PD or HD. They might also have to switch between dialysis types as a result of of high blood pressure, fluid overload or infections.

Hence, it is no wonder that discussions on all three options make up a considerable amount of our corpus. In the opening posts, the synset {haemodialysis, hemodialysis} is mentioned 785 times, and the synset {organ transplant, transplant, transplantation} even appears in a total of 1,409 documents. Table 2 displays the list of synsets most correlated with the respective treatments. To provide some perspective: the average correlation coefficient of the pairs formed by the 1,876 distinct synsets in the corpus above the threshold level is approximately 0.0435 with a standard deviation of 0.0598.

Table 2: Table displaying synsets correlated with treatment options. CIs are adjusted to control the false coverage-statement rate (FCR) at level 5%. To save space, each synset is identified by one representative synonym, e.g. {dark, night, nighttime} is displayed as {night}.

<table>
<thead>
<tr>
<th>Synset</th>
<th>Corr</th>
<th>95% CI</th>
<th>Synset</th>
<th>Corr</th>
<th>95% CI</th>
<th>Synset</th>
<th>Corr</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>donor</td>
<td>0.34</td>
<td>0.32, 0.36</td>
<td>peritoneal</td>
<td>0.25</td>
<td>0.23, 0.27</td>
<td>catheter</td>
<td>0.28</td>
<td>0.26, 0.30</td>
</tr>
<tr>
<td>kidney</td>
<td>0.33</td>
<td>0.31, 0.34</td>
<td>intervention</td>
<td>0.18</td>
<td>0.16, 0.20</td>
<td>exchanged</td>
<td>0.19</td>
<td>0.17, 0.21</td>
</tr>
<tr>
<td>time period</td>
<td>0.28</td>
<td>0.26, 0.3</td>
<td>patient</td>
<td>0.16</td>
<td>0.14, 0.18</td>
<td>time period</td>
<td>0.16</td>
<td>0.14, 0.18</td>
</tr>
<tr>
<td>list</td>
<td>0.24</td>
<td>0.22, 0.26</td>
<td>home</td>
<td>0.15</td>
<td>0.13, 0.17</td>
<td>way</td>
<td>0.15</td>
<td>0.13, 0.17</td>
</tr>
<tr>
<td>recipient</td>
<td>0.24</td>
<td>0.22, 0.26</td>
<td>discipline</td>
<td>0.13</td>
<td>0.11, 0.15</td>
<td>begin</td>
<td>0.14</td>
<td>0.12, 0.17</td>
</tr>
<tr>
<td>rejection</td>
<td>0.23</td>
<td>0.21, 0.25</td>
<td>modality</td>
<td>0.13</td>
<td>0.11, 0.15</td>
<td>nurse</td>
<td>0.14</td>
<td>0.12, 0.16</td>
</tr>
<tr>
<td>wait</td>
<td>0.23</td>
<td>0.21, 0.25</td>
<td>nephrology</td>
<td>0.12</td>
<td>0.09, 0.14</td>
<td>bag</td>
<td>0.12</td>
<td>0.10, 0.14</td>
</tr>
<tr>
<td>message</td>
<td>0.22</td>
<td>0.2, 0.24</td>
<td>change</td>
<td>0.11</td>
<td>0.10, 0.13</td>
<td>commutation</td>
<td>0.12</td>
<td>0.10, 0.14</td>
</tr>
<tr>
<td>years</td>
<td>0.21</td>
<td>0.2, 0.23</td>
<td>occurrence</td>
<td>0.11</td>
<td>0.09, 0.13</td>
<td>dialysis</td>
<td>0.11</td>
<td>0.08, 0.13</td>
</tr>
<tr>
<td>experience</td>
<td>0.21</td>
<td>0.19, 0.23</td>
<td>experience</td>
<td>0.11</td>
<td>0.09, 0.13</td>
<td>home</td>
<td>0.11</td>
<td>0.08, 0.13</td>
</tr>
<tr>
<td>donate</td>
<td>0.21</td>
<td>0.19, 0.23</td>
<td>therapy</td>
<td>0.11</td>
<td>0.09, 0.13</td>
<td>night</td>
<td>0.11</td>
<td>0.08, 0.13</td>
</tr>
<tr>
<td>information</td>
<td>0.21</td>
<td>0.19, 0.23</td>
<td>death</td>
<td>0.11</td>
<td>0.09, 0.13</td>
<td>weaken</td>
<td>0.10</td>
<td>0.08, 0.13</td>
</tr>
<tr>
<td>create</td>
<td>0.21</td>
<td>0.19, 0.23</td>
<td>failure</td>
<td>0.11</td>
<td>0.09, 0.13</td>
<td>fluid</td>
<td>0.10</td>
<td>0.08, 0.13</td>
</tr>
<tr>
<td>transfer</td>
<td>0.20</td>
<td>0.19, 0.22</td>
<td>rate</td>
<td>0.10</td>
<td>0.08, 0.12</td>
<td>feel</td>
<td>0.10</td>
<td>0.08, 0.12</td>
</tr>
<tr>
<td>possession</td>
<td>0.20</td>
<td>0.18, 0.22</td>
<td>years</td>
<td>0.10</td>
<td>0.08, 0.12</td>
<td>peritoneal</td>
<td>0.10</td>
<td>0.07, 0.12</td>
</tr>
</tbody>
</table>

Whereas forum members mentioning transplants seem to be most interested in discussing the modalities of receiving a kidney transplant from a donor, a large amount of discussions on PD and HD seem to involve comparisons between the two options, e.g. the most highly correlated synset with {haemodialysis, hemodialysis} is {peritoneal}. And indeed, 10% of discussions on the two treatments mention both HD and PD.

A commonly raised objection concerning HD is the need to visit a dialysis center, with many patients preferring to undergo dialysis at home, either in the form of home HD or PD. {home} frequently co-occurs with both PD and HD. PD patients or those considering it seem to be mostly discussing its practical aspects: Synsets like {catheter}, {exchange}, {bag} and {dark, night, nighttime} all refer to the immediate process of PD. Because of this practical
bias in the results, we assumed that a large proportion of users discussing PD have practical experience. Manually checking a random sample of 100 opening posts mentioning PD, we find that 67 did already undergo PD treatment while 33 did not, so significantly more patients belong to the former group (95%-CI: 0.57, 0.76).

Besides considering the three different treatment options, some of our other observations are the following:

1. We found a disproportionate occurrence of the synset \{hubby, husband, married man\} compared to \{married woman, wife\}, with only 122 first threads mentioning the latter and 433 mentioning husbands.
2. Of all the 3,174 first threads, 617 reference an ancestor, most of the time either mother or father.
3. A very large number of discussions centered around food, with \{food, nutrient\} appearing in 2,623 of 10,216 opening posts. Besides several ingredients, we found \{formula, recipe\} to be highly correlated with the food synset (Corr: 0.195, 95% CI: [0.176, 0.215]).

The first observation could mean that women constitute the majority of forum members or at least of those who care for a significant other. To check this, we manually labelled the gender of 200 users by looking at their opening threads. We positively identified 94 members, of which 67 were women and 27 men. The 95% CI for the percentage of female users is (0.58, 0.76), indicating that significantly more women than men are active in the considered CKD discussion forums. Because of the second observation, we might guess that up to 20% of all opening threads were created by someone caring for a relative. Out of a random sample of 100 threads, we have manually identified 79 posts coming from someone caring for an ancestor, in most cases father or mother, and five threads alluding to a question on the hereditary nature of the disease. 16 threads were of an unspecified nature. This confirms our hypothesis that offspring of patients are concerned about the health of their parents and seek help online. Notice that in both cases, we restricted our view to the first threads created by each user. We decided to do this as it ensures that heavy users do not bias the results when inferring a characteristic of the forum population. The third observation triggered the hypothesis that while patients may in theory know what their diet should look like, they could be missing recipes that can help them to practically adapt to the new dietary requirements. In the documents, \{formula, recipe\} appeared 220 times, and in a random sample of 100 of opening posts containing both food and recipe, 40 posters specifically asked for recipes, 36 shared recipes and 24 discussed some other recipe-related topic. These results suggest that 3% of users who discussed food specifically asked for recipes. (95%-CI: 0.026, 0.042).

5 Discussion

We found Internet self-help forums to be an unique and potentially fruitful medium for learning about the daily issues CKD patients face. Our results indicate that there might be a need for stronger embedding of relatives in the treatment process. We also detected a gender gap, with significantly more women active on the forums than men. Further research might provide insights whether this finding applies only to the two considered forums or generally, and if so, what measures could be taken to increase men’s participation. A bit surprising was our observation that many people seem to discuss alternatives to classical hemodialysis such as home hemodialysis or peritoneal dialysis, even though the percentage of patients on these treatments combined is only around 10% in the US. It is unclear to us whether the clientele opting for these more unpopular treatment options is particularly contracted in the two considered forums or whether in general there exists a much larger demand for PD and home hemodialysis. Another explanation for this might be that patients engaged in these forums are highly motivated and more active in their disease management. Finally, a significant number of patients inquired about the types of food they were still allowed to eat, asking for recipes satisfying their nutrition requirements.

As we have seen, self-help forums accumulate a lot of information over time. This makes the forums data hubs that patients can query to answer their questions. However, not all of the accumulated information is accurate, and the shared first-hand experiences might not be representative. Therefore, it is necessary to investigate the potential biases existing in the opinions and experiences expressed on these forums. In addition, analyzing the differences in patient’s reactions towards first-hand experiences of fellow patients compared to the advice of their doctors might shed further light on how patients’ needs are addressed best.

Analyzing online communication data comes with a lot of challenges, requiring the development of new approaches for text analysis. The results of our proposed methodology are encouraging, but there are several potential improvements.
One of them would be to enhance the disambiguation algorithm, which is an active research topic in itself. One limitation of our current algorithm is that the similarity score for words of different parts of speech is always zero due to the WordNet hierarchy. Hence, if there was only one word of a given part of speech in a context, the most frequent synset is chosen without taking the neighboring words into account. A potential remedy could be to leverage WordNet’s synset definitions in the similarity calculation. This would be similar to the Lesk algorithm\textsuperscript{34} which was adapted to WordNet previously.\textsuperscript{35}

Alternatively, disambiguation could be improved by extending the Brown corpus or by integrating ontologies such as UMLS to compensate WordNet’s deficiencies in the medical domain. Another limitation of our current method is that it does not take syntactic information into account and thus relies on a human-guided final step. If this could be incorporated, one would not only pertain semantic information, but could also identify the interactional relations: Who acts and who is acted upon? However, given the complexity of such a task, this might be a long-term project. From a short-term perspective, it might be more fruitful to develop methods for calculating synset similarity and clustering techniques of synset trees in order to more easily identify discussed topics. Since our methodology is not tied to a single medical domain, one could think of applying it in other areas. In particular, we think of forums where due to the nature of the problem one would expect high veracity (chronic or lethal diseases) or where due to a high social stigma there is a lack of reliable information (e.g. depression).

6 Conclusion

Given the persistent rise in costs related to CKD, developing new preventive measures to curb disease progression has become an economic necessity. In this study, we have developed a process to systematically analyze online self-help forums. We encountered a few surprising phenomena: The large gap in user activity, the gender gap, forums as knowledge hubs and the desire for practical information. The study of online self-help forums broadens the perspective and allows medical practitioners to gain insights into the daily routine of chronic disease patients. Incorporating the insights drawn from such analyses might allow doctors to take a more holistic view of their patients and promote more personalized treatments. Information extraction and analyses of communication of chronic disease patients might therefore play an important role in developing new preventive care approaches.

Acknowledgement

We wish to thank David Choi for his contributions and valuable feedback provided in many discussions about the project.

References


Dynamic Estimation of the Probability of Patient Readmission to the ICU using Electronic Medical Records

Karla Caballero, PhD Candidate\(^1\) and Ram Akella, PhD\(^{1,2}\)

\(^1\)University of California Santa Cruz, 1156 High Street Santa Cruz CA, USA
\(^2\)University of California Berkeley, 94720 Berkeley CA, USA

Abstract

In this paper, we propose a framework to dynamically estimate the probability that a patient is readmitted after he is discharged from the ICU and transferred to a lower level care. We model this probability as a latent state which evolves over time using Dynamical Linear Models (DLM). We use as an input a combination of numerical and text features obtained from the patient Electronic Medical Records (EMRs). We process the text from the EMRs to capture different diseases, symptoms and treatments by means of noun phrases and ontologies. We also capture the global context of each text entry using Statistical Topic Models. We fill out the missing values using a Expectation Maximization based method (EM). Experimental results show that our method outperforms other methods in the literature terms of AUC, sensitivity and specificity. In addition, we show that the combination of different features (numerical and text) increases the prediction performance of the proposed approach.

Introduction

Currently, the accurate and opportune prediction of patient readmission to the ICU shortly after he is transferred to a lower level care is of great interest to health providers. Tools that estimate the probability of patient relapse and readmission to the ICU aid physicians and health care providers to determine the possible resources that should be allocated to the patient and to discover the possible causes of relapse that could lead to the patient to be readmitted to the ICU. The timely and accurate estimation of patient’s probability of readmission allows us to successfully trigger a medical alarm before the patient is transferred from the ICU.

This probability estimation also permits the early identification of patients with elevated risk of readmission. As a result health care providers can differentiate those patients from the ones who are stable and less likely to return in order to assign medical resources more effectively.

Most of the existent methods in the literature\(^{1,2}\) rely on the use of static classifiers that do not take into account the evolution of the patient or the dynamic nature of the patient’s features. To overcome these challenges, we propose a dynamic method based on Bayesian Time Series and Dynamic Linear Models (DLM) to estimate the probability of readmission before the patient is discharged to indicate the existence of a possible medical alarm. Our contribution is summarized as follows: We model the probability of patient readmission as an aggregated latent state which is updated each time new features are observed (lab results, vital signals, etc.). Our model is fed with heterogeneous data obtained from his Electronic Medical Records (EMRs) which consists of text and numerical data with both discrete and continuous variables. We incorporate the text information into the model by developing a method which converts the unstructured text information into discriminative features that are later incorporated into the model. Finally, we address the missing values problem by estimating those values using a Regularized Expectation Maximization (EM) based method.

In this context, we find that the dynamic estimation of the probability of patient readmission to the ICU provides a more accurate prediction when compared with other methods in the literature and other static prediction methods.
Background

Prevailing medical practice relies on frameworks such as the Apache III\textsuperscript{3}, and SAPS II\textsuperscript{4} scores. Both methods, widely used to predict patient mortality in the Intensive Care Unit (ICU) are used as proxy to estimate the likelihood that a patient is not ready to be transferred to a lower level care and there is a high probability of being readmitted if he is discharged from the ICU. These methods incorporate temporal information in a limited way by only choosing the worst-case scenario values during the first 24-hour window that a patient is inside the ICU. As a result, they often overestimate the probability of mortality (not readiness to be transferred from the ICU). Moreover, these scores are only estimated once during the entire stay in the ICU, which may not indicate whether the patient would recover and be successfully transferred from the ICU in the future.

Data Mining has been previously used to address the problem of estimating the likelihood that a patient is readmitted to the ICU\textsuperscript{1,2}. Other prediction methods such as Batal et al.\textsuperscript{5} include the dynamic information by collapsing the time series of features, such as blood pressure and heart rate, into static features that are later used in a classification framework. However, this model does not take into account the evolution of the patient in time. Most of the existing solutions rely on training a static classifier with a patient’s observed feature vector. These features are mainly static, such as lab reports produced at a given point of time or the estimated APACHE score at the discharge time.

Furthermore, most of the methods mentioned above assume the availability of all the features at the prediction time. This assumption may not be valid in a real scenario where the data is often incomplete and segmented. Health care data suffers from a large volume of missing data due to the fact that not all the features are collected (lab results, vital signal, etc) for all the patients at all time. One of the most common methods to fill out these missing values is to perform mean imputation. However, this practice has been shown to introduce more noise into the model rather than reduce it\textsuperscript{6}. To tackle this problem, previous approaches segment the patient features according to their age group and then calculate the average value for each segment\textsuperscript{7,8}. Other methods handle missing values by fitting a distribution for each feature with the observed data and sample from the estimated distribution when the value is missing\textsuperscript{9}. Similarly, the use of Multiple Imputation to predict the missing values has been proposed previously. Here regression techniques with the other observed features as covariates are deployed\textsuperscript{6,7}. Overall, these methods do not take into account the temporal aspect of the missing data where current features values are highly dependent on previous values.

Most of the existing prediction models do not use text from the Electronic Medical Record (EMRs) due to its complexity. However, text data contains key information that is potentially useful to better predict the likelihood that a patient is readmitted to the ICU. Examples of text include lab reports, admission, doctors and nurse notes. Ghassemi et. al\textsuperscript{10} combine static numerical features such as SAPS II score with topic modeling features from the text of the EMRs to estimate the probability that a patient die after 30 days of being discharged using Support Vector Machines (SVM).

Our proposed approach combines text and numerical information in a dynamic setting that allows us to predict the patient readmission before the patient is transferred from the ICU to other hospital areas. In addition, our proposed approach takes into account the evolution of the patient health state and temporal aspect of the patient features to predict how likely he will relapse and will be readmitted to the ICU.

Methods

In this section, we describe the method to construct the probability of readmission as a latent state, the methodology we use to extract numerical features and to process the text information to extract the discriminative features. In addition, we outline the framework we use to handle missing values. Finally, we describe the experimental settings and the methodology to validate the proposed approach.

Definition of Probability of Readmission as a Latent State

We define the that a patient \(i\) would be readmitted to the ICU in the next 30 days \(Y_{t,i}\) or not \(Y_{t,i} = -1\) if it is discharged from the ICU at time \(t\) as a binary variable. \(Y_{t,i}\) is 1 with a probability of \(\pi_{t,i}\), which represents the probability that a this patient is readmitted if he is transferred from the ICU at time \(t\). (probability of readmission). This probability is a function of a latent state \(\theta_{t,i} = [\xi_{t,i}, \tilde{\theta}_{t,i}].\) This latent state is formed by the estimation of the log-odds of the probability of readmission in previous steps \(\xi_{t-1,i}\) and the patient features observed at time \(t, \tilde{\theta}_{t,i}\). The value of \(\tilde{\theta}_{t,i}\) is obtained by combining a set of observed features \(X_{t,i}\) obtained from the EMR at time \(t\) and the value
of this combination at previous steps, $\hat{\theta}_{t-1,i}$.

In this framework, we are able to include both the patient’s features and his health context obtained from previous time steps. This is not accounted for in the static classification frameworks. Our proposed model is a special case of the Generalized Dynamic Linear Models (GDLM)\(^{11}\). Here, we employ the logistic transformation to accommodate our specific context. This leads to the following expressions:

\[
Y_{t,i} \sim \text{Bernoulli}(\pi_{t,i}), \quad \pi_{t,i} = \frac{e^{\xi_{t,i}}}{1 + e^{\xi_{t,i}}}, \quad (1)
\]

\[
\xi_{t,i} = \xi_{t-1,i} + \tilde{\theta}_{t,i} + w_{t,i}, \quad w_{t,i} \sim N(0, W_{\xi}) \quad (2)
\]

\[
\tilde{\theta}_{t,i} = \lambda \tilde{\theta}_{t-1,i} + \beta X_{t,i} + w_{t,i}, \quad w_{t,i} \sim N(0, W_{\theta}) \quad (3)
\]

Here $\lambda$ is a decay factor that determines the contribution of previous feature values in the current value of $\tilde{\theta}_{t,i}$. The vector $X_{t,i}$ is constructed from the patient’s observed lab test results, vital signals readings, text notes features, and demographics (features). In this model we assume that most of the values of $X_{t,i}$ are observed. In later subsections of this paper, we explain how we model and estimate the missing values of the feature vector. The vector $\beta$ represents the regression coefficients we use to combine the observed features. The value of $\theta_{t,i}$ can take both positive and negative values. Thus, we are able to increase or decrease the probability of readmission using the observed features $X_{t,i}$. $W_{\xi}$ and $W_{\theta}$ are the evolution variances of $\xi$ and $\theta$ respectively.

The value of $\xi_t$ reflects the log-odds effect on the probability of readmission $\pi_{t,i}$ by previous observed features contained in the state $\theta_{t,i}$. To illustrate this effect, we calculate the impact of the user’s features $X_{t,i}$ observed at time $t$ and then aggregate them into the state $\theta_{t,i}$ after $k$ steps assuming no other values of $[X_{t+1,i} \ldots X_{t+k,i}]$ are observed. This impact is determined by the following forecast function:

\[
\xi_{t+k} = \sum_{r=0}^{k} \lambda^r \tilde{\theta}_r = \tilde{\theta}_t (1 - \lambda^{k+1}) / (1 - \lambda) \quad (4)
\]

As illustrated by the previous equation, the proposed model incorporates knowledge from prior measurements into the current state estimation. This effect representation allows us to predict patient probability of readmission even when no measurements are available at a given time $t+k$. In addition, the effect does not decrease over time, as opposed to $\theta_t$ (observed features). Each time there are new observations available, the value of the effect $\xi_{t,i}$ is updated using equation 4.

**Model Fitting**

Figure 1 shows the graphical model of this framework. The colored circles represent the variables that are observed in the model. The non-colored circles are the latent variables and model parameters that need to be inferred. The learning across multiple users is reflected through the estimated parameters $\Phi$ defined as: $\Phi = \lambda, W_{\theta}, W_{\xi}, \beta$. This representation is flexible enough to expand the model and to incorporate different weighting vectors $\beta$ for different patient groups with a particular disease or age range.

We fit the model using Dynamic Linear Models with Logistic Transformation. This model incorporates the user features into an aggregated patient state that evolves over time, in contrast to static classification models. In addition, we train the model using the entire patient’s stay path as opposed to individual time steps. By performing this, we take into account the uncertainty about the future in the estimation of the probability of readmission.

Our proposed approach allows us to predict future values of the state as more readings become available. Consequently, we are able to dynamically estimate the current patient probability of readmission and predict its evolution using the predictive forecast function of the latent state. Figure 2 describes the fitting steps for the proposed model.

In the following subsections, we describe the process to extract both numerical and text features, as well as the method to fill out missing values. Once we have the time series features for all the patients in the training set, we give an initial value to the model parameters $\Phi$.

We train the model using an iterative method based on Expectation Maximization (EM). This method consists of 2 steps: E and M steps. In the E step, we estimate the latent state of the patient $i$, $\theta_{t,i}$ using the Forward Filtering
Backward Smoothing method (FFBS)\textsuperscript{12}. In this method, we estimate first the latent state $\theta_{t,i}$ using the values of the observed features $X_{t,i}$ and the state value of the previous time step $\theta_{t-1,i}$ (Forward Filtering). Once we have estimated the entire path, we correct the estimated latent state backwards using the estimated state values of future time steps $\theta_{t+1,i}$ (Backward Smoothing). By combining the Forward Filtering (FF) and the Backward Smoothing (BS), we guarantee the construction of a fully dynamic model with feedback where previous values of the path affect the current latent state while accounting for the future uncertainty. One variant of this dynamic model is to train the model with no feedback about the future (open loop feedback). To achieve this, we fit the model using the latent states obtained with the Forward Filtering (FF) step only.

The M-step consists of estimating the values of the parameters $\Phi$ that optimize the latent patient paths (probability of readmission path) previously estimated in the E-step. Then, we repeat the E and the M steps until convergence.

We predict the probability of readmission in the test partition assuming that we do not know the final outcome of the patient. Therefore, we estimate the latent state $\theta_{t,i}$ using the estimated parameters and previous values of the latent state $\theta_{t-1,i}$ using Forward Filtering.

**Numerical Features Extraction**

We extract the numerical features using the events described in the EMR. A patient may be subject to different procedures and events during his stay in the ICU based on his condition. The events that we incorporate into the model are selected by means of the $\chi^2$ test. We estimate then this score for all the events that appear in the corpus and retain those with higher score (the most discriminative ones). We extract 30 features such as blood pressure level, lab procedures, pain level, and heart rate.
We observe that the selected features are a combination of those used by the APACHE III and SAPS II scores. We find that 80% of the Apache III score features and 90% of the SAPS II features are included in our selected features. This selection shows consistency between our proposed approach and these widely used methods of the literature.

We identify that some of the selected features are considered to have a bimodal distribution. For instance, it is equally dangerous to have a really low blood pressure as to have it to be really high. To integrate this knowledge, we assign a weight for each possible range of the event. Those weights are obtained from the those used by the Apache Score III.

We then divide the selected user features into two groups: quasi-static and dynamic. Some of the labs and procedures do not need to be performed at each time step. Therefore, we consider these features as quasi-static (they are updated if there is a new reading). Features such as blood pressure, pain level and heart rate are considered to be dynamic. This division impacts how missing values are treated in these features. The quasi-static features are updated when a new value is observed. Meanwhile, the dynamic features will be filled in using the method to estimate missing values described later in the paper. In addition to these features, we include two demographic features about the patient: gender and age which we treat as static variables. Both features provide us the initial conditions to set the initial probability of readmission. Table 1 shows a summary of the feature extraction process.

Text Feature Extraction

Standard approaches used as proxy to estimate the probability of readmission, such as Apache III and SAPS II scores, do not incorporate text information. One of the main challenges researchers face is to incorporate this type of data effectively and to create discriminative features.

To achieve this, we extract text features to improve the health state prediction, consequently the estimation of the probability of readmission to the ICU. The text entries found in an EMR mainly consist of nurse’s entries, procedures reports, admission and discharge information, among others. Each text entry has an assigned timestamp. Thus we are able to construct a time series for each of the text features we extract. In this subsection, we describe the steps we follow to process the text and extract different features that are later integrated into the model. Figure 3 depicts the text-based feature extraction process.

Noun Phrases Extraction

Given the nature of health care text data, we need to extract meaningful phrases and concepts to successfully represent diseases and treatments (features) which help us to improve our statistical estimates. To achieve this task, we extract noun phrases relevant to the medical domain by annotating a set of discharge summaries using the Clinical Text Analysis and Knowledge Extraction System (cTAKES) and Metamap. With these tools we extract clinical named entities (concepts) such as drugs, diseases/disorders, signs/symptoms, anatomical sites and procedures.

Discharge Summaries provide us a rich data set to explore and to extract noun phrases. These documents often aggregate the patient’s medical history. This includes all the patient’s information collected during his stay in the ICU, treatments and care information that the patient should follow after he is discharged from the ICU.

After extracting all the entities, we select the phrases which describe a disease, a procedure or a medication using the medical ontologies provided by SNOMED (matched concepts). In addition, we also detect which set of noun...
phrases corresponds to stop words (i.e. patient name, doctor name). By means of tf-idf term selection, we select the most important noun phrases and remove those with low score.

Once the phrase selection is completed, we perform standard stop words removal and stemming before indexing the documents. Then, we extract two types of features: term and topic based features which we describe below.

**Term-Based Features**

We incorporate into the model a term-based feature using the obtained noun phrases and word terms from the EMR text entry. This feature is based on the classification of the text entries: $-1$ if the patient is ready to be discharged and $1$ if not. We use the Naive Bayes classifier, which has been shown to provide good predictive performance and it is computationally feasible for the unbalanced classes problem\textsuperscript{16}.

In order to make this classification feasible, we reduce the large vocabulary size of the corpus by extracting the most discriminative terms by means of $\chi^2$ test of the probability of readmission\textsuperscript{17}. We performed this test on the whole corpus. Our goal is to obtain a global estimate from the whole corpus in order to reduce the bias resulting from the term selection. We keep 4000 terms after this step.

**Topic-Based Features**

The second set of text features is based on statistical topic modeling. These models allow us to reduce the dimensionality of the term space to a smaller feature space of latent ”topics”. In addition, we are able to model topics for unseen documents without training the model again, as the method is generative.

In this context, each document is represented as a mixture of topics with a certain probability. Similarly, each topic is represented as a mixture of words. Our hypothesis is that topics capture the global context of the document while this cannot be achieved by selecting text terms alone. By capturing this context, we are able to improve the performance of the prediction of patient readmission.

For this application, we fit a GD-LDA model to extract the topics from the corpus set using 75 topics. The corpus consists in all the processed text entry notes (noun phrases + terms) of all the patients. GD-LDA, which is a generalization of LDA\textsuperscript{18}, allows us to model correlations between topics as opposed to LDA. In addition, this method is fitted in an unsupervised form, and it is computationally efficient, which permits us to train a large number of documents in a single batch contrary to other statistical topic models that model correlations such as Correlated Topic Models (CTM)\textsuperscript{19}.

We then remove the background topics which we define as word mixtures with a high percentage of common words (more than 90% of the terms inside the topic). We define as common words those that do not have healthcare related information by comparing them with the ontologies from the UMLS using MetaMap\textsuperscript{14}. These ontologies provide
information about healthcare treatments, drugs and diseases. We keep 65 topics after this step.

After removing the background topics, we select the 10 most discriminative topics by means of the $\chi^2$ test and include them in the dynamic model as features. Our hypothesis is that patients with certain topics in their EMR are more susceptible to be readmitted to the ICU than those whose EMR does not have them. In order to make the documents comparable, we use the values of \{1, 0\} to show the presence or absence of a topic in the document instead of the probability of the topic in the document (two patients with similar medical history can have the same topics in their EMRs, but in different proportions). Thus, we indicate that a topic is present in a document if it accounts for more than 5% of the total topic mixture inside the document. In addition to the most discriminative topics, we include the classification output of the text entry (patient ready to be discharged or not) using the document topic mixture as features. Here we use Naive Bayes classifier.

**Missing Features Estimation**

The proposed framework mentioned above assumes that most of the patient’s features are observed at each point in time. When feature values are not observed, we indicate they are missing and then we impute their value.

For the current application, the patient’s features have an implicit temporal aspect. The value of features in time $t$ are highly dependent on their value in previous time steps $1..t-1$. Then, standard imputation methods based on the mean value could lead to estimation errors. To overcome this challenge, we impute the missing values by means of a Regularized Expectation Maximization method. This iterative method uses the observed features at a particular time to impute the missing ones using an initial value of the parameters (E-step). Then we choose parameters that minimize the error rate between the observed values and the imputed ones (M-step).

The E-step (expectation) consists of estimating the missing features $\hat{x}_m$ using the values of the available ones $x_a$ inside the record using the following equation:

$$\hat{x}_m = \mu_m + (x_a - \mu_a)\gamma, \quad \gamma \sim N(0, \Sigma) \quad (5)$$

where $\mu_m$ is the mean estimate of the missing features of a record and $\mu_a$ is the mean estimate for the available features in a given record. We define as a record all the observed features of a patient at a given point in time $X_{t,i}$. The value of $\gamma$ is the vector of regression coefficients for the available features. The hypothesis behind this step is to represent the missing record values as a combination of the available values for a given record and an estimated mean of the
missing values of all the records.

After the missing values have been imputed, we perform the M-step. In this step, we estimate the mean and variance of the observed features and coefficient weights of how these are combined to impute a missing value. We repeat both steps until a certain error rate is achieved (around 5%). More details of the algorithm can be found in the work of Schneider.20

Experimental Settings and Validation

We test our approach by estimating the probability that a patient discharged from the ICU and transferred to other areas of the hospital would be readmitted in the near future. This prediction is critically important in the accurate prognosis of patient health state (we want to avoid health complications in the patient by discharging him before time). In addition, this prediction is very important to assign medical resources effectively.

For the situation which we are modeling, we predict the probability that the patient would be readmitted if he is discharged at time \( t \) using the information available in his EMR before he is discharged. The EMRs are obtained from the MIMIC II data set.21 This dataset contains text and numerical information that describes procedures, medications and vital signs readings from a given patient during his stay in the ICU. MIMIC II is composed of medical records from over 30,000 patients admitted to the ICU during a 7 year window of hospitals from Boston Area.

We validate our method using the EMRs of 15,000 patients selected randomly. In order to compare our approach with other methods from the literature, we only study the adult patients (over 18 years of age) without excluding patients due to a specific illness; this data consists of 11,648 people. We report our results using five-fold cross validation, taking 80% of the patients as training set and the remaining 20% as test set. We construct the time series of each patient by aggregating the patient information every 3 hours.

We report the probability of patient readmission after \( t = 24 \) and 48 hours and at the time of discharge. Our goal is to test the prediction capability of the proposed approach at different time steps. We compare our method with related methods of the literature used as proxy to estimate the probability of readmission such as Apache Score3 and SAPS II4. In addition to these methods, we compare the proposed approach with static classification methods such as Naive Bayes and Random Forests with 50 trees with a sliding window for \( t = 24, 48 \) and before patient discharge as in our proposed method.

Results

When analyzing the time series of the patients, we detect that the 57% of the patients stayed 24 hours of less. Thus common practices used as proxy to detect the probability of patient readmission such as Apache Score and SAPS II are not applicable for those patients. We also detect that after dividing the features in quasi-static and dynamic features 34% of the features are not observed. Therefore, the method to fill out missing values would have high impact in the estimation of the results.

Figure 4 shows the obtained topics after constructing noun phrases and removing background topics. Note that by performing these steps we are able to obtain cleaner topics, each one related to a particular disease describing symptoms and body parts than those provided by Ghassemi et al10. In addition to the quality of the topics, we also evaluate the quality of the topic-based classification output using the document topic mixture as features. Here we observe that 72% of the notes from the discharged patients that were readmitted were classified correctly and 52% of the notes from the patients that were not readmitted to the ICU were classified correctly using topic-based features. In addition, we evaluate the performance of the classification using term-based features. Here we observe that 53% of the notes from readmitted patients were predicted correctly and 10% of the notes from patients non readmitted were classified correctly. When comparing both features, we observe that topic-based classification is more accurate than term-based classification. In this context, we corroborate our hypothesis that topics capture the global context of the of the text notes and improve the classification performance when compared with term-based classification.

Table 2 shows the AUC, sensitivity and specificity for the proposed method and other methods proposed in the literature. Here we observe that our proposed framework outperforms the other reported methods in the three reported

\(^1\)K-fold cross validation provides a more robust prediction evaluation than leave-one-patient-out-method because K-fold cross validation uses less training data
measures and the time steps reported. We notice that in the other methods their prediction performance decreases after 24 hours. Compared to these methods, our proposed approach does not decrease the performance significantly after 24 hours of patient admission. The reason behind this fact is that the evolution of the latent state and the compilation of the feature data from previous time steps into the current one.

In addition, our proposed approach has a good balance between sensitivity and specificity. A high specificity value implies the existence of a small number of false alarms. In a real scenario, this measure has a high impact due to the limited medical resources that health providers have. Physicians do not want to be overloaded with false alarms at the time a true alarm arrives. But at the same time, they need to be predicting accurately the existence of a true alarm. Detecting all true alarms correctly is desirable since the cost of not detecting a patient who is not ready to be discharged who is very ill and dies outside the ICU is very high.

In table 2, we also observe that mortality prediction models such as Apache III and SAPS II provide a good proxy to estimate the probability of readmission. However the main limitation of these methods is that they tend to overestimate the probability of readmission as we can observe in the specificity measure. The best static prediction method is based on Random Forests at $t = 24$ hours.

Note that the combination of topic, text based and numerical features provide the best prediction performance. Therefore, we corroborate our hypothesis that text data provides complementary information to numerical features and increases the performance in the prediction of patient readmission. When analyzing the average probability of readmission, in our model with all the features, we notice that this probability increases from 0.69 in $t = 24$ hours to 0.89 in before patient discharge to those who were not readmitted compared to 0.90 in $t = 24$ hours to 0.92 before the patient is discharged from those who were readmitted to the ICU. This shows that the stay length affects the estimation of the probability of patient readmission. This phenomenon can also explain the performance decrease of the other prediction methods.

**Discussion**

In this paper we have proposed a dynamic model that combines heterogeneous data form the patient Electronic Medical Records to predict the patient readmission to the ICU. The accurate prediction of patient readmission would lead to an good allocation of resources that potentially could reduce cost in the transfer and patient care.

The use of an aggregated patient state which combines current features with previously observed ones allows us to predict the probability of readmission even if not new features are observed. The results of the proposed model depend on the quality and quantity of the patient observed features. The more features values are observed for a given patient in his EMR and the more diverse patient pool data is, the more accurate readmission prediction would be.

The current model provides a single aggregated latent state. Future work includes expanding the model to create of different latent states, one latent state for each body subsystem. Our final goal is to obtain a global prediction estimate.

| Table 2: Performance of the 3 variations of our model, 2 different methods used in the literature and 2 static classification methods in terms of Sensitivity, Specificity and AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Apache III | 0.9258 | 0.4042 | 0.8665 | 0.8812 | 0.1289 | 0.5320 | 0.9294 | 0.8006 | 0.8574 | 0.9016 | 0.2435 | 0.7263 |
| SAPS II | 0.8707 | 0.4280 | 0.8119 | 0.8707 | 0.1289 | 0.5320 | 0.9294 | 0.8006 | 0.8574 | 0.9016 | 0.2435 | 0.7263 |
| Naive Bayes | 0.8812 | 0.1289 | 0.5320 | 0.8694 | 0.8116 | 0.8110 | 0.8694 | 0.8116 | 0.8110 | 0.8694 | 0.8116 | 0.8110 |
| Random Forests | 0.9294 | 0.8006 | 0.8574 | 0.928 | 0.2247 | 0.7587 | 0.928 | 0.2247 | 0.7587 | 0.928 | 0.2247 | 0.7587 |
| Proposed method | 0.8094 | 0.8116 | 0.8110 | 0.8870 | 0.5069 | 0.8094 | 0.8984 | 0.8445 | 0.8202 |
| Numerical Features | 0.8692 | 0.9244 | 0.9070 | 0.9087 | 0.8494 | 0.8429 | 0.9138 | 0.9378 | 0.9412 |
| Proposed method | 0.9043 | 0.8833 | 0.9289 | 0.9138 | 0.9378 | 0.9412 | 0.9149 | 0.8964 | 0.9274 |
| Text and Numerical Features | 0.9043 | 0.8833 | 0.9289 | 0.9138 | 0.9378 | 0.9412 | 0.9149 | 0.8964 | 0.9274 |
| Proposed method | 0.9043 | 0.8833 | 0.9289 | 0.9138 | 0.9378 | 0.9412 | 0.9149 | 0.8964 | 0.9274 |
| Topic, Text and Numerical Features | 0.9043 | 0.8833 | 0.9289 | 0.9138 | 0.9378 | 0.9412 | 0.9149 | 0.8964 | 0.9274 |

In 24 hours compared to these methods, our proposed approach does not decrease the performance significantly after 24 hours of patient admission. W e notice that in the other methods their prediction performance decreases after 24 hours. Compared to these methods, our proposed approach does not decrease the performance significantly after 24 hours of patient admission. The reason behind this fact is that the evolution of the latent state and the compilation of the feature data from previous time steps into the current one.
which results from the combination of the subsystems latent states. In the same manner, we plan to expand the model to accommodate different feature weight vectors, one for each disease group. Our hypothesis is that feature values are different across multiple diseases.

Acknowledgments

This work was partially supported by the NIST grant number 60NANB13D136, by NSF/NIST/UMBC grant number SC-0000015277, by CONACYT grant number 207751, by CITRIS SFP 2011-164, and by CITRIS SFP 2015-325. Authors would like to thank Ashit Talukder for his inputs.

References

11 West M, Harrison J. Bayesian forecasting and dynamic models (2nd ed.). Springer-Verlag; 1997.
Modelling Risk of Cardio-Respiratory Instability as a Heterogeneous Process

Lujie Chen, MS1,2, Artur Dubrawski, PhD2, Gilles Clermont, MD3, Marilyn Hravnak, PhD4, Michael R. Pinsky, MD5
1Heinz College and 2School of Computer Science, Carnegie Mellon University, and School of Medicine3 and School of Nursing4, University of Pittsburgh, Pittsburgh, PA

Abstract
Cardio-respiratory instability (CRI) occurs frequently in acutely ill. If not identified and treated early, it leads to significant morbidity and mortality. Current practice primarily relies on vigilance of the clinical personnel for early recognition of CRI. Given limited monitoring resources available in critical care environment, it can be suboptimal. Thus, an “Early Warning Scoring” mechanism is desirable to alert medical team when a patient is approaching instability. It is widely recognized that critically ill may show subtle changes prior to the onset of CRI, but it is not well known how their risk evolves before the onset. Using large amounts of physiological data routinely gathered from continuous noninvasive monitoring of Step-Down Unit patients, we demonstrate a data-driven approach that: (1) Characterizes patient’s individual CRI risk process; (2) Identifies groups of patients that progress along similar risk evolution trajectories; (3) Utilizes grouping information to help forecast the emergence of CRI.

Introduction
Cardio-respiratory instability (CRI) is commonly observed in critically ill patients. If not caught early, CRI may escalate and require costly interventions and result in poor outcomes1. In current practice, prompt recognition of CRI often heavily relies on the vigilance and experience of the clinical staff2, which presents a challenge especially in environments where case load is high and experienced staff is in shortage3. The study by Hravnak et al.2 shows that about 25% of the patients may develop at least one episode of CRI during their Step-Down Unit(SDU) stay, however, about 83% of CRI episodes occurred without being noticed, and among those that were actually recognized, the average latency of detection may be as long as 6.3 hours.

Given that a large amount of multi-parameter continuous vital sign measurements are readily available from existing bedside monitors commonly used in critical and post-operative care, it is not only desirable, but also possible to develop “Early Warning Scoring” (EWS) systems and incorporate them into the current monitoring process that is primarily driven by human judgment. In our previous work4, we developed such a scoring model to predict CRI. Similarly to other EWS frameworks, it can estimate with some accuracy “if” and “when” a patient might deteriorate, however, it offers little insight into “how rapidly” it could happen. Knowing the likely rate of crisis escalation in individual patients would be tremendously valuable for clinical resource allocation and it could substantially improve medical outcomes.

We propose a data-driven approach that builds upon physiologic monitoring data to discover phenotypes of CRI risk evolution during periods before its onset. It takes a “bottom-up” approach to paint a microscopic view of each individual patient’s risk process towards CRI by explicitly modeling its dynamics as well as the heterogeneity observed between the individuals. To estimate the instantaneous risk score, it uses a popular supervised classification approach. Temporal sequences of the estimated risk scores are then used to learn a finite mixture model of risk score trajectories that are approximated as third order polynomials. The mixture model groups the individual risk trajectories into a parsimonious number of equivalence classes: the risk trajectory groups. Patients who are grouped together share similar apparent dynamics of their risk evolution in the period preceding the onset of CRI, while the risk evolution patterns differ between distinct groups.

In the first part of the paper we describe the approach we take to model risk of CRI and its evolution, and how the individual risk trajectories are grouped into phenotypic equivalence classes. We then proceed to demonstrate how the ability to identify assignment of a patient to its most likely phenotype group could help forecast the onset of CRI, especially in patients whose risk escalation is rapid and who are therefore at the highest chance of being missed under the current critical care practice until their condition substantially deteriorates. We illustrate our proposed modeling methodology with empirical evaluation on real-world bed-side monitoring data collected in a Step-Down Unit (SDU) of a hospital that includes 532 annotated monitoring episodes that escalated to a real CRI episode, combined with 370 controls.
This work has relevance to some of the EWS literature. There are basically two categories of EWS. The first type is the triage EWS. Often computed just once, at the time of admission, triage risk scores are static and designed to communicate the patient’s long-term risk in general (such as e.g. mortality) without specifying a time horizon of the possible occurrence of any particular future adverse events. Example of a triage EWS is the well-known APACHE score\(^5\) or its newer alternatives that use advanced machine learning models\(^6\). Our work however is more closely related to the second type of EWS that is designed for continuous monitoring purposes. Rothman Index\(^7\) is an example. This index is computed from 26 variables extracted from the electronic health record that include vital signs, nursing assessments, laboratory test results and cardiac rhythms, and it communicates the risk of mortality within the next 24 hours, among other outcomes. Rothman Index can be updated at regular intervals (most typically hourly) or when new information about the patient’s status becomes available. Though not part of its design, if sufficiently informative data is available, scores like Rothman Index could potentially be used to model trajectories of risk evolution. In that sense, our work can be thought of as a “meta-analysis” of a particular continuous updated bedside monitoring score.

In the setting of continuously monitored patients, our modeling framework is similar to what Guiza et al.\(^10\) developed for Neuro-ICU patients to predict the Increased Intracranial Pressure (ICP) which is a risk factor for poor outcome. They built a single model to predict the ICP events to occur in 30 minutes based on dynamic features extracted from a 4-hour period of observation in a mixed cohort of positive and control patients. In our previous work\(^4\), we instead built 30 independent models to predict CRI at subsequent one-minute intervals starting 30 minutes before the hypothetical onset of the event. This makes it possible to reason about the temporal trends in CRI predictions using the data available at any given point in time, which laid the basis for our current investigation.

Our approach is also related to the concept of dynamic risk process. Wiens et al.\(^11\) estimated daily risk of C.diff colonization in hospitalized patients through a classifier that contrasted positive patients (who eventually tested positive) and control patients (those who never tested positive, or have never been tested) using a large number of features extracted from electronic health records as inputs. They reported a time series of Area Under the Receiver Operating Characteristic Curve (AUC) scores within days before the C.diff events to characterize the temporal pattern of the overall model performance. We have explored a similar approach to modeling temporal patterns of model performance for CRI events in our previous work.\(^4\) In this paper, however, we take a step further to analyze the underlying longitudinal risk process for each event for each individual patient. This allows us to have a detailed view of the risk evolution trajectories beyond what can be obtained by using the cross-sectional aggregated temporal AUC scores.

**Methods**

**Data**

Non-invasive continuous vital signals were collected from bedside monitors for patients in a 24 bed surgical/medical trauma Step-Down Units in a large urban teaching hospital during a three month period. Heart rate (HR), respiratory rate (RR) and peripheral oxygen saturation (SpO\(_2\)) were recorded at frequency of 1/20 Hz (3 readings per minute), blood pressures measurement (SysBP, DiaBP) were recorded intermittently with the minimum frequency of once every 2 hours.

**Identification of CRI events**

CRI events often occur in patients who show obviously abnormal vital signs for an extended period of time, the onset of CRI may signify the switch of patient’s state from being relatively “stable” to “unstable” and thus require clinical intervention. In order to identify real CRI events, and separate them from clinically irrelevant artifact often present in bed-side monitoring data, we first identify vital sign events (VSEs) which are the individual instances of any of the vital signs exceeding a predefined threshold\(^8\). We then apply temporal persistence criterias to filter out VSEs that last less than 3 minutes and/or the duty cycle of the observed exceedance is lower than 80%. Guided by the CRI event annotation protocol\(^9\), a committee of 6 clinical experts adjudicated the resultant sets of events in order to discriminate true CRI events from the artifacts. The artifact events refer to the events manifesting in data that have no relevance to patient’s physiological state, such as side effects of body movement, monitoring hardware or software malfunctions, etc. We only used the events adjudicated as true CRI in our analysis described below.

**Inclusion and exclusion criteria**

We have included all patients with least one CRI episode during their SDU stay in our analysis. If one patient had multiple events, all could be included, with the exception of the events that occurred within one hour of admission
(for the first CRI event of a given patient) or if they occurred within one hour after the ending of a prior event (for subsequent CRI events recorded during the particular hospital stay). Since the goal of our work is to model the patient’s risk of progression towards a CRI from their currently stable state, these data filtering criteria ensure that the pre-event periods are indeed “event free” and that the patients during these time periods can be considered clinically stable.

The patients who have never experienced any CRI during their stay became our control group. For them, we identified hypothetical event onset timestamps during their first 6 hours of the SDU stay. The timestamps were spaced evenly 45 minutes apart unless there were no sufficient data available. To be more conservative about our control set, we opted to use only the data from the initial periods of SDU stay, since most patients tend to stabilize further as their stay at SDU extends.

Feature extraction

In order to analyze the dynamics of the risk processes, for each event, either positive CRI or control, we identified 30 reference timestamps, equally spaced at 1-minute intervals ending at the onset and starting up to 30 minutes before the CRI onset time. Together with the onset time (lag=0), each patient/event combination had 31 sets of observations at those subsequent timestamps, or less if no sufficient data was available. For each observation, a set of numerical features was derived from the vital signs during the time windows of 5, 10, and 15 minutes, ending at the particular timestamp. Simple statistical features of vital sign time-series were derived separately for each vital (HR, RR, SpO2, SysBP, DiaBP), including mean, variance, min, max, range, linear slope, and fraction of valid readings, for each of the three window-width settings. This resulted in a total of 105 (3 window sizes x 5 vitals x 7 statistics) features for each observation (i.e. a patient/event/timestamp combination). Fig. 1 illustrates the feature extraction process.

![Feature extraction framework](image)

Figure 1. Feature extraction framework.

Modeling the risk process of as related to an individual event

We derived the risk process for each patient/event combination using a supervised machine-learning framework. We first constructed a training set of observations from positive and control data described above, with each observation represented by an array of 105 features derived from the interval of 15 minutes immediately preceding the onset of the events (i.e. lag=0), either true CRI or a control non-event, and we assigned a binary label as either positive or negative, accordingly to the output dimension of each such data point. We then trained a random forest classification model from this data. We chose the random forest approach for illustrative purposes only. It is likely other popular classification models could be used in its stead with comparable success. Our choice was partially motivated by positive personal experience with this type of classifier in various application domains and due to the ability of the particular implementation used to gracefully handle missing data.

The trained random forest model was then used to score the data observed during periods preceding the onset of CRI or control non-events, separately and independently for each of the 30 reference timestamps, and for each of the patient/event combination from a held out data set that was not used to train the classifier to mitigate the risk of over-fitting.

The predictions made by the trained classifier on the held-out data served as the estimated risk scores. Their values range from 0.0 to 1.0, with the higher values reflective of a greater similarity of the currently observed vital signs
features to those of patients going into an episode of CRI, and conversely, lower values suggest resemblance to the non-events obtained from control patients. Computing these scores over time at one-minute intervals, we obtained time series of scores that can be used to reflect the risk process as it is leading towards the event of interest.

We conducted our experiments using a 10-fold cross validation protocol, where the complete set of patients was split into 10 disjoint subsets. In each of 10 iterations, one of these subsets was used as a test set and the concatenation of the remaining nine subsets as the training sets. We split the data in a particular way to make sure that (1) the same patient’s data would belong to either training or a test set, but never to both in the same cross-validation iteration; (2) the proportion of positive and negative observations stayed relatively similar across all cross-validation iterations.

Group models of risk processes

We then analyzed the collection of the individual risk processes derived in the previous step, using the Group Based Modeling (GBM) approach13. In this approach, temporal patterns of patients’ risk processes are modeled as a mixture of polynomial functions with timestamps as the covariates. For a given risk trajectory of length T, assuming K groups, and the likelihood function of the trajectory is provided as follows:

$$\sum_{i=1}^{K} \pi_i \prod_{j=1}^{T} P_i(Y_j | X_j)$$

in which \(\pi_i\) and \(P_i\) is the prior probability and probability function respectively for group indexed by \(i\). \(Y_j\) is the response variable value at timestamp indexed at \(j\), in our case it is the estimated risk score from random forest classifier; \(X_j\) is the covariate, in our case it is the timestamp value (i.e. \(X_j = j\)).

Given the assumed number of distinct groups of risk trajectories (K) and the chosen order of the polynomials, the model estimates smooth trajectories for each of the groups characterized by the shape of the polynomial functions. The probability function form is determined by the response variable, in our case, since the response variable is a risk score bounded between 0 and 1, we used the truncated normal distribution. The model also estimates the prior distribution of the group membership across modeled data. The model parameters are inferred using Maximum Likelihood method through a numerical optimization procedure. We used the SAS procedure “traj” described in this paper14 and available from this website15.

After the GBM model is trained, posterior group membership for a currently observed risk process trajectory of an individual can be computed on-the-fly. Knowing the likely membership of a patient to one of the previously identified risk trajectory classes, we can inform the medical personnel of the likely near-term evolution of this patient’s status towards a possible CRI episode. The posterior membership is computed as follows:

$$P(group = i | Y, X) = \frac{\hat{\pi}_i \prod_{j=1}^{T} \hat{P}_i(Y_j | X_j)}{\sum_{i=1}^{K} \hat{\pi}_i \prod_{j=1}^{T} \hat{P}_i(Y_j | X_j)}$$

Prediction algorithm

We used the trained group model above as a part of the prediction algorithm designed to estimate the expected patient’s risk status \(t\) minutes in the future from now. In order to study the utility of the risk process grouping information for forecasting CRI, we set up two alternative protocols that took the grouping information into consideration.

In the first of these protocols, the feature set of a random forest classifier included the previously described 105 statistical features of the vital signs, but also the maximum posterior group membership indicator as an additional discrete feature. In the training phase, group membership is obtained from the fully observed individual risk process (i.e. “now casting” trajectories), while in test phase it is estimated from an online prediction of the group membership based on the currently observed vital sign features and the resulting estimated risk scores. We call these current snippets of risk score time series “trajets” since they are smallish segments of a longer risk trajectory.

The second risk-process-group aware protocol also used the current 105 statistical vital sign features, but each risk trajectory class was treated with a separate random forest risk estimation model trained using only the data of patients who have been assigned to the particular group. This was equivalent to enforcing that each decision tree in
the random forest predictor used the group label as the root node feature, while the former approach allowed the grouping information to be leveraged anywhere in the decision tree hierarchies.

As a baseline, we included the third protocol that did not use estimated groupings to forecast CRI at all. It however used the 105 time series features only, consistently with what we have been doing in our previous work.

In testing, the current patient’s data is processed using the trained risk score estimation models to obtain a “trajlet”: a short time-series of estimated risk scores. The trajlet is then processed by the trained GBM model to predict the risk process group category this trajlet most likely belongs to. The group information is then combined with the 105 features extracted to make a prediction. In the third, baseline protocol, the prediction is simply obtained by feeding the 105 features together with the group information into the single trained random forest classifier. The second protocol follows that but including predicted trajectory membership information as the additional input feature. In the third protocol, since separate models are built for each group, we only make the prediction for the most likely group assignment, and to that we use the model trained for that group and the 105 statistical features.

We empirically evaluated performance of those algorithms at predicting CRI to set on in 20 minutes. We used AUC as the performance metric. Figure 2 depicts the information flow for the first and the second protocol.

Figure 2. Information flow for the first and second protocol used to forecast CRI.

**Results**

**Data for experiments**

Subject to the inclusion/exclusion criteria described above, we included 158 positive patients and 71 control patients in our analysis, from which 532 positive events and 370 non-events were identified. The average number of events is 3.4 for each positive patient (CRI) and 5.2 for each control patient (non-events). For each event, 31 observations were generated at timestamps of 0, 1, 2,… and up to 30 minutes before the CRI onset time. This amounted to 25,983 observations. Demographic information of the positive and negative cohorts is presented in Table 1. The right hand side of the table lists the prevalence of admission diagnosis category and medical history conditions deemed to have possible relevance to CRI occurrences. As expected, control patients had significantly shorter lengths of SDU stay than the target group; however they did not reveal significant differences in any other observed characteristic.

**Estimating risk process trajectories and risk process grouping**

A total of 902 individual risk evolution sequences (532 positives and 370 negatives) were generated using the supervised learning method described above. We present here the result from a 10-fold cross validation experiment in which the risk estimates are pooled from 10 disjoint test sets in the 10-fold setup. A finite mixture model of truncated normal distribution was used with the GBM approach. We chose to use K=5 groups, each modeled with a polynomial of third order.
Table 1. Patient demographics and CRI related admission diagnosis categories and medical history conditions.

<table>
<thead>
<tr>
<th>Patients</th>
<th>Positive</th>
<th>Control</th>
<th>p-value</th>
<th>Admission Diagnosis Categories</th>
<th>Positive</th>
<th>Control</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>158</td>
<td>71</td>
<td></td>
<td>Circulatory</td>
<td>28</td>
<td>23</td>
<td>0.105</td>
</tr>
<tr>
<td>% male</td>
<td>55%</td>
<td>61%</td>
<td>0.564</td>
<td>Digestive</td>
<td>26</td>
<td>7</td>
<td>0.477</td>
</tr>
<tr>
<td>age (mean, years)</td>
<td>60.35</td>
<td>55.36</td>
<td>0.079</td>
<td>Injury</td>
<td>96</td>
<td>39</td>
<td>0.754</td>
</tr>
<tr>
<td>Charlson Deyo Index (mean)</td>
<td>0.98</td>
<td>0.83</td>
<td>0.406</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDU length of stay (mean, hours)</td>
<td>127.02</td>
<td>66.18</td>
<td>&lt;0.001</td>
<td>Myocardial Infarction (MI)</td>
<td>17</td>
<td>2</td>
<td>0.067</td>
</tr>
<tr>
<td>Monitoring</td>
<td></td>
<td></td>
<td></td>
<td>Congestive Heart Failure (CHF)</td>
<td>16</td>
<td>6</td>
<td>0.811</td>
</tr>
<tr>
<td>Total monitoring hours</td>
<td>3,930</td>
<td>17,532</td>
<td></td>
<td>Chronic Pulmonary Disease (CPD)</td>
<td>31</td>
<td>8</td>
<td>0.132</td>
</tr>
<tr>
<td>Mean monitoring hours</td>
<td>111</td>
<td>55</td>
<td>&lt;0.001</td>
<td>Cerebral Vascular Disease (CVD)</td>
<td>13</td>
<td>7</td>
<td>0.801</td>
</tr>
</tbody>
</table>

Figure 3 presents the 5 groups of risk process trajectories learned from our data. Each summarizes CRI risk progression for the cross validated set of patients during 30 minutes before the onset of their CRI event (or non-event for the controls). Solid lines in Figure 3 show the representative risk process trajectories for each group aggregated over the individual risk trajectories underlying the particular group, where the group assignment is based on their maximum posterior probability of group membership given full length of trajectories. Dotted lines depict means and confidence intervals estimated from the trajectory group models. On the horizontal axis, timestamp 0 indicates the onset of CRI. Vertical axis scale shows the risk scores. The graph also includes group size information (with count on left and proportion in parenthesis) obtained from the GBM model. Table 2 presents the distribution of labels and types of CRI events (stratified by the type of vital sign that was first to cross the control limits) for each group. Figure 4 shows the most likely trajectories for each of the group based risk process models, overlaid with the corresponding individual risk processes, each smoothed with third order polynomials.

As can be seen, those trajectories reveal five distinct CRI risk trends. The largest group is group 2 (trajectory plotted in yellow, “persistently low risk”), including 43% of the test cases with about 2/3 of them coming from control patients. The characteristic of this group is a flat trend of low risk over time. As can be noticed in Figure 4, individual patient trajectories that belong to this group also include some positive cases with very late (that is, rapid) onset of CRI. Most of the time these patients look stable, but then they suddenly develop CRI without leaving much lead time for the early warning system to recognize the change. With these exceptions aside, patients who could be classified as members of this group are not likely to require medical attention in the near future.

Group 5 (pink, “persistently high risk”) includes about 18% of all cases, with about 94% of them associated with positive events. This group is characterized by a persistently high risk of CRI. These patients have been in poor
condition before our observation started and they remain in it throughout. Medical personnel is likely already aware of their need for close scrutiny, and early warnings are of little use in these cases.

Group 1 (red, “late onset”) and group 4 (blue, “early onset”) contain mostly positives cases (82% positives for group 1 and 90% for group 4). Different from the other groups, they exhibit gradually increasing trends of risk, though at different lead times of the onset of the patient’s status change. Group 4 picks up the risk escalation trend as early as about 25 minutes before the CRI event, while group 1’s trend looks very similar to that of group 2 (which has a negative majority) until about 10 minutes before CRI onset. These groups are very interesting from the clinical perspective. If we could establish for a patient with the apparently low risk at the moment, that their likelihood of becoming unstable would escalate in the near term, and in addition to that we could establish with some certainty how rapid will the escalation be, this would provide medical personnel with very valuable information about the outlook of all individuals under their care so that they could prepare, prioritize, and take preemptive actions if possible, to mitigate the consequences of the imminent health status deteriorations.

Lastly, group 3 (green, “transient”) is comprised of about equal number of positive and negative cases, and it is the hardest to classify apart from others. Unless a patient can be confidently assigned to the “persistently low risk” group, they will require close monitoring, so do patients belonging to this particular group.

These results show potential utility of the proposed approach to modeling evolution of risk in managing critical care resource allocation. It also has a potential to favorably impact patient outcomes by characterizing the likely rate of escalation of the imminent crisis in a particular patient, prompting medical personnel to act with a desired urgency.

In column (B) of Table 2, we show the fraction of all positive cases can be reliably identified at various lead times as members of groups 1, 4 or 5. The combined recall of patients who will eventually become unstable across these three groups is about 59% with a specificity of 92% (false positive rate of 8%). In columns (C) and (D) of Table 2, we show the true positive (TP) and true negative (TN) counts that would be produced if we were to accurately predict group membership which can be used to predict the outcome (CRI vs. non-event) using the class majority vote rule. The recall rate of such system would be 77% at a specificity of 72% (false positive rate of 28%).

Considering the distribution of data and results according to which vital sign was first exceeded during an episode of CRI, we notice a relatively small portion of BP and HR events, and when they occur they often seem to be hard to separate from negative cases. About half of RR events belong to the “persistently high risk” group, while 17% of them belong to the “late onset” group and about 16% to the “early onset” group. For SpO2 events, 30% are mixed with negatives in the “persistently low risk” group but they are most likely the cases with the late sudden onset. Additional 27% of SpO2 events group with “late onset” class and 12% with the “early onset” class, while 17% belong to the “persistently high risk” group.

Forecasting future occurrences of instability

In this section, we report preliminary experimental results of an attempt to use risk trajectory group estimation to support forecasting of CRI. So in this case, we only observe a portion of the full risk trajectory (a “trajlet”), from which we can make online determination of the group membership. Then, we can leverage the predicted grouping to predict expected patient’s CRI risk changes in the near time.

In this experiment, we set the current time to 20 minutes before the event. This leaves us with the 10 initial minutes of observations to form a “trajlet” that becomes the input for group classification algorithm. This can be done on-the-fly using the group model learned from training data. In Figure 5, AUC scores of three risk estimation models described above split by the estimated group label. Model #1 (red) is the single random forest model with an additional discrete feature encoding predicted group membership; Model #2 (green) uses five independent random forest models, one for each risk trajectory group, and Model #3 (blue) is the baseline single random forest model with no group information incorporated. In Figure 6, we present ROC curves for each of these models, separately for each CRI risk evolution trajectory group.
Table 2. Profiles of the identified risk process groups.

| Group ID | Type                   | Lead time | % of Pos | BP  | HR  | RR  | SpO₂ | Pos(%) | Neg(%) | Total | TP    | TN    | (A)  | (B)  | (C)  | (D)  |
|----------|------------------------|-----------|----------|-----|-----|-----|------|--------|--------|-------|-------|------|------|------|------|
| 1        | late onset             | ~ 10 min  | 19%      | 2   | 9   | 38  | 53   | 102(82%)| 23(18%)| 125   | 102   | 56   | 52   | 224  | 200  |
| 2        | persistently low risk  |           |          | 30  | 18  | 18  | 59   | 125(32%)| 266(68%)| 391   | 266   | 158  | 94   | 158  | 94   |
| 3        | transient              | ~25 min   | 12%      | 0   | 2   | 36  | 24   | 62(90%)| 7(10%) | 69    | 62    | 94   | 41   | 158  | 104  |
| 4        | early onset            |           |          |     |     |     |      |        |        |       |       |      |      |      |      |
| 5        | persistently high risk | >30 min   | 28%      | 4   | 8   | 104 | 33   | 149(94%)| 10(6%) | 159   | 149   | 56   | 52   | 224  | 200  |
| Totals   |                        |           | 59%      | 56  | 52  | 224 | 200  | 532    | 370    | 902   | 407   | 266  |

Figure 4. Group trajectories overlaid with smoothed individual risk process trajectories participating in these groups.

We observe some divergence of performance between the baseline model and group based models, primarily in the context of group 1. This is the “late onset” group. It is interesting to see that group classification information available at 20 minutes before the onset of CRI, when the current risk score of group 1 does not differ much from the risk score of the “persistently low risk” group 2, is helping boost forecasting ability about the eventual outcome when compared to the model that does not use group assignment information. This suggests that group information carries some predictive value which is not captured effectively by the numeric features used by the baseline random forest model. This opens possibility for effective use of group information to inform clinical personnel of imminent health crises even before the moment when momentary online risk scores begin to escalate (20 minutes before the onset vs. approximately 10 minutes for group 1).

For group 2 (“persistently low risk”), all three models show very similar performance. This is not surprising since these patients often show consistency throughout the period of observation and their risk trends can be easily captured by random forest classifier even without using the group information. In case of group 4 (“early onset”), the risk trend is about to pick up around the time of making the prediction (20 minutes before the event), and it seems that the baseline model was already able to capture the information needed for forecasting, hence group information also does not help much here. It should be noted that it should be possible to repeat the interesting result discussed in the context of group 1 above, if we moved the time of observation further back. Group 5 (“persistently high risk”) is primarily composed of positive cases, with very limited representation of negatives, and the lack of
class balance makes ROC characteristics look unimpressive for either of the three considered models. It is also noted that the integrated performance (five groups combined) compared across three models shows no significant difference. This is understandable since this result is driven by the differences in performance between groups and the relative group sizes. In our case, groups 2, 3 and 4 where performance variation is small, account for over 80% of the cases in data. Given the integrated performance that is similar across the three models, at the false positive rate of 30%, sensitivity is 69%, positive predictive value is 76% and specificity is 70%.

Figure 5. AUC scores compared among three models for each group as well as for all 5 groups combined, showing 95% empirical bootstrap confidence intervals.

Figure 6. ROCs for each group and each model.

Conclusion

We described a framework designed to model evolution of risk as a heterogeneous process with the application to tracking emergence of Cardio-Respiratory Instability using continuous vital signal available from bedside monitors. We have evaluated this framework empirically using representative cohort of step-down unit patients. The presented framework is intuitive and it allows us to model explicitly the dynamics of risk changes for individual episode for individual patient. This new capability has a potentially useful clinical application in helping to understand how the patients deteriorate and not only if they will, and if so, when.

Using a group based modeling approach we discovered distinct types of risk evolution processes. Some patients exhibit persistently high-risk status, some show persistently low risk status, and some fit models that follow particular trajectories of risk escalation. The former type of patients either already receive attention of the clinical personnel, or they are not in urgent need to receive it, but patients who could be reliably identified as one of those who are on a rapid trajectory leading towards a crisis, could be flagged by the presented method for immediate attention of clinical personnel. These quickly escalating instances are particularly hard to capture by the current monitoring protocols.

Our framework establishes a linkage between the risk process and the predictive performance of the models. The distinct types of trend patterns suggest various degrees of predictability in terms of lead time and reliability. Assuming the trend patterns reflect the true underlying risk progress of the patient group, then the predictability revealed from the grouping represents an upper bound of the performance that an early warning score algorithm may achieve. It helps explaining variations of performance at various lead times and among different cohorts drawn from the population. The common practice of reporting an aggregated model performance for entire population at a fixed
point of time may mask the rich underlying dynamics and heterogeneity across the individual patients and their phenotypic equivalence classes that could be informative and clinically relevant. In this aspect, our framework is general and it can be directly adapted to any other modeling scenario involving temporal escalations or de-escalations towards or away from adverse or favorable events in healthcare and beyond.

In the context of clinical care, our method reside in between of the population-level analysis where a universal model is built for the whole population of interest, and a personalized approach where individual characteristics of the individual patient are the only source of predictive information. Both these approaches have their limitations, most prominently the risk of over-generalization in the population-level modeling case, and the risk of over-specification and the perils of low data supply in the personalized modeling case. Our framework lies in between of those two extremes allowing for heterogeneity between the individuals but also providing data support to make reliable predictions possible by drawing the required evidence from the patient’s group peers.

Additional utility of our work is in providing explanatory capability by visualizing current and expected future risk trajectory and updating it dynamically at frequent intervals, in order to provide salient and timely information to the medical personnel in an intuitive and understandable fashion.

Acknowledgements

This work has been partially supported by NIH (R01NR013912) and NSF (1320347).

References

Combining Human Disease Genetics and Mouse Model Phenotypes towards Drug Repositioning for Parkinson’s disease

Yang Chen, MS¹, Xiaoshu Cai, BS¹, Rong Xu, PhD²

¹Department of Electrical Engineering and Computer Science, School of Engineering, 2Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA

Abstract

Parkinson’s disease (PD) is a severe neurodegenerative disorder without effective treatments. Here, we present a novel drug repositioning approach to predict new drugs for PD leveraging both disease genetics and large amounts of mouse model phenotypes. First, we identified PD-specific mouse phenotypes using well-studied human disease genes. Then we searched all FDA-approved drugs for candidates that share similar mouse phenotype profiles with PD. We demonstrated the validity of our approach using drugs that have been approved for PD: 10 approved PD drugs were ranked within top 10% among 1197 candidates. In predicting novel PD drugs, our approach achieved a mean average precision of 0.24, which is significantly higher (p<e-11) than 0.16 for a state-of-art drug discovery approach based on mouse phenotype data. Comparison of gene expression profiles between PD and top-ranked drug candidates indicates that quetiapine has the potential to treat PD.

Introduction

Parkinson’s disease (PD) is the second most common neurodegenerative disorder and affects 5 million people throughout the world¹. The prevalence of PD is expected to grow significantly with the aging of population². Current pharmacological treatment options are primarily dopamine replacement strategies³,⁴. Among them, levodopa remains the most effective agent in treating PD³, but may cause serious adverse events, such as motor and neuropsychiatric complications⁵. In addition, the dopaminergic therapies usually fail in reversing the progressive neuronal loss and controlling non-dopaminergic features of PD, such as sleep impairment, dementia, and sensory disturbances, which have become a major source of disability for most patients²,³,⁴. Therefore, new drugs are desired to improve the treatment of PD.

Computational drug repositioning approaches lead to rapid drug discovery. Most current repositioning strategies can be categorized as drug-based, disease-based, and profile-based (Figure 1). Drug-based (Figure 1(a)) and disease-based (Figure 1(b)) approaches exploit drug-drug or disease-disease similarity and existing drug-treatment knowledge to infer new disease-drug associations⁶,⁷,⁸. Drug similarities are based on pharmacological data, such as drug chemical structure⁹ and drug side effects¹⁰. Disease similarities are calculated using disease phenotype⁷, disease genetic and genomic data¹¹. Recently, profile-based repositioning strategies (Figure 1(c)) have successfully found new drugs for inflammatory bowel disease¹², small cell lung cancer¹³, and lung adenocarcinoma¹⁴. These approaches do not rely on the knowledge of existing drug treatments, and have increased ability to discover new drug-disease pairs compared with drug-based and disease-based strategies.

Current profile-based drug repositioning approaches mainly exploit the gene expression profiles of drugs and diseases¹²,¹³,¹⁴, and have an inherent challenge: the lower-level genomics profile similarities between drugs and diseases do not necessarily translate into higher-level drug treatment efficacy in diseases. Recently, the Mouse Genome Informatics (MGI) has made available large amounts of phenotypic descriptions for mouse genetic mutations based on systematic gene knockouts¹⁵, which are impossible on human. These causal gene-phenotype associations in mice have been demonstrated useful in discovering of new disease genes¹⁶,¹⁷ and drug targets¹⁸, and also have the potential to overcome the challenge in profile-based drug repositioning approaches. Hoehndorf and colleagues have used mouse phenotypes in their drug repositioning approach¹⁹. They first predict disease genes through comparing human and animal phenotype ontologies, and then link the predicted disease-gene associations with the drug-target data to suggest new drug indications. In this study, we present a novel profile-based drug repositioning strategy to exploit both lower-level disease and drug genetics and higher-level mouse phenotypes. We apply the approach for PD drug discovery and compare with Hoehndorf’s approach.
Our approach leverages disease-gene associations, drug-target relationships, and mouse phenotype data to predict new drugs for PD. We first identified PD-specific mouse phenotypes using human disease genes in OMIM, and then screened all the FDA-approved drugs for candidates that share similar mouse phenotype profiles with PD. Recent studies have demonstrated that disease-associated genes have potential to guide drug discovery. Mendelian genetics in Online Mendelian Inheritance in Man (OMIM) has been underappreciated in drug target identification, but in fact provides invaluable opportunities for drug repurposing. To date, OMIM has included 15 high-penetrance PD genes that are likely to cause the PD symptoms in the mice carrying their mutations. Though these genes are mostly associated with familial PD, clinical studies have shown that the familial and sporadic forms of PD usually share the same molecular pathways. We compared our approach with the pure genetics-based approaches and demonstrated the importance of mouse phenotypes in improving the performance of PD drug detection. We also compared with Hoehndorf’s approach and show that combining disease genetics with mouse phenotypes using our novel approach achieves significantly better precision. Finally, we analyzed the gene expression profiles for top-ranked candidate drugs and the result suggests that quetiapine has the potential treat PD.

Figure 1. Computational drug repositioning strategies: (a) disease-based methods, (b) drug-based methods, and (c) profile-based methods.

Figure 2. Drug discovery approach for Parkinson’s disease combining human disease genetics and mouse mutation phenotypes.
Methods

Our hypothesis is that a drug has the potential to treat PD if the drug target genes are associated with PD phenotypes in mice. Gene-phenotype associations based on systematic mouse gene knockouts provide rich information to link drugs and their new indications. Figure 2 shows that our drug repositioning approach based on mouse phenotypes contains two steps. In the first step, we searched for the mouse phenotypes associated with PD using the well-studied disease genes. In the second step, we extracted a set of mouse phenotype features for each candidate drug and systematically calculated the semantic similarities (using mammalian phenotype ontology) of the phenotype profiles between PD and candidate drugs. Using the mouse phenotype similarity between the drugs and disease, we predicted how likely the drugs can be used to treat PD.

A. Identify mouse model phenotypes for PD using disease genetics in OMIM

We searched for mouse model phenotypes for PD using 15 genes associated with 20 subtypes of PD in OMIM. The mutations of these genes highly increase the risk for PD and are likely to cause PD phenotypes. All these human genes have homologies among mice. We downloaded the phenotype annotations for mouse genes from MGI, and extracted 358 phenotypes that are linked to the 15 PD genes. Different PD genes may share common phenotype annotations. For example, 7 out of 15 PD genes point to the phenotype of neurodegeneration. We weighted each phenotype with the number of its associated PD genes. The weights intuitively represent the confidence that the phenotype is related with PD.

We ranked the PD-specific mouse model phenotypes by their weights, and investigated the category of the top-ranked phenotypes. The mammalian phenotype ontology classifies mouse phenotypes into 30 categories. We first mapped each PD phenotype to its categories by tracing the isa relationship in the mammalian phenotype ontology. The 358 phenotypes were mapped into 24 categories. Then we calculated a score for each category by summing the weights of all the phenotypes in it. We ranked the categories based on these scores and examined the top five categories.

B. Prioritize candidate PD drugs based on the similarities of mouse phenotype profiles between disease and drugs

We collected a set of candidate drugs from DrugBank. The drug-target database in DrugBank contains information for 1427 FDA-approved (for any indication) drugs. We extracted 1197 drugs that target on human/mouse orthologous genes, and included them into the candidate drug set. Then we combined the drug-target relationships and phenotype annotations for the target genes to link each candidate drug to a set of mouse model phenotypes through the drug target genes. We constructed a vector of mouse phenotypes for each drug, and weighted each phenotype by the number of its associated target genes.

We calculated the semantic similarity between the vector of mouse phenotypes associated with PD and each candidate drug to determine how likely the drug can be used to treat PD. A similar calculation of semantic similarity between ontology concept vectors was used in a previous study. We first quantified the information content for each phenotype term as \(- \log p(t)\), in which \(p(t)\) represents the frequency among phenotype annotations to all the 7568 mouse genes. In calculating the information content, if a gene is annotated by one phenotype term, we assumed that it is also annotated by the ancestors of this term in the hierarchy of mammalian phenotype ontology. Hence, a phenotype term has higher information content than its ancestors, which lie on higher levels in the ontology. Then we defined the semantic distance as:

\[
sim(t_1, t_2) = \max_{a \in A(t_1, t_2)} \log p(a),
\]

where \(A(t_1, t_2)\) is the set of common ancestors for \(t_1\) and \(t_2\). To calculate the distance from the phenotype vector \(p_1\) to \(p_2\), we matched each phenotype term in \(p_1\) to the most similar term in \(p_2\) and took the average:

\[
sim(p_1 \rightarrow p_2) = \frac{1}{|A(t_1, t_2)|} \sum_{t_2 \in p_2} \max_{t_1 \in p_1} \sim(t_1, t_2).
\]

The matching similarity was weighted by the product of weights for phenotype term \(t_1\) and \(t_2\). The similarity between \(p_1\) and \(p_2\) was defined as the average of semantic similarities in both directions:

\[
sim(p_1, p_2) = \frac{1}{2} \sim(p_1 \rightarrow p_2) + \frac{1}{2} \sim(p_2 \rightarrow p_1).
\]

C. De novo evaluation in prioritizing FDA-approved PD drugs

We investigated if our method can prioritize approved PD drugs. We ranked the 1197 candidate drugs using the semantic similarities of the mouse phenotype profiles between the drugs and PD. Then we extracted FDA-approved
PD drugs from FDA drug labels. Our drug ranking algorithm does not use any information of the approved PD drugs. In the de novo evaluation, we calculated the distribution of approved PD drugs among our ranks by plotting a 10-bin histogram. Specifically, we divided the ranks into 10 ranges, and counted the number of approved PD drugs within each range. In addition, we investigated the target genes for the top 10% candidate drugs. We ranked these drug target genes by the number of drugs (ranked within top 10%) that target on each gene. We also calculated the distribution of genes targeted by the FDA-approved drugs among all the drug target genes using histogram.

We demonstrated the importance of using mouse phenotypes to predict drugs for PD. Recent studies have shown that disease-associated genes can guide the detection of existing drug therapies and promising candidate drugs.22,23 We compared our approach with two genetics-based drug discovery methods (figure 3). The first method directly matches the disease genes in OMIM with the drug target genes to repositioning existing drugs for new indications. The second method extends the disease-associated genes with their functionally related genes based on protein-protein interactions (PPIs), and matched the extended gene set with the drug target genes for drug repositioning. We downloaded the PPIs from the STRING database, and used the experiment data source, which contains PPI databases such as HPRD, BIND, and GRID. We evaluated if the two methods have the ability to identify approved PD drugs without using mouse phenotypes, and compared the result with our approach.

![Figure 3](image)

**Figure 3.** Our approach compared the mouse phenotype features between disease and drugs. We compared with genetics-based drug discovery methods, which directly match the disease genes and their interacting genes with the drug target genes. The comparison aims at demonstrating the importance of using mouse phenotypes.

D. Evaluation in ranking novel PD drugs and comparison with an existing drug repositioning approach

We investigated if our approach has the ability to prioritize novel PD drugs. In our recent studies, we constructed large-scale drug-disease treatment knowledge bases from multiple data resources using techniques including natural language processing, text mining and data mining.22,33,34 The databases included 9,216 drug-disease treatment pairs extracted from FDA drug labels, 34,306 pairs extracted from 22 million published biomedical literature abstracts, and 69,724 pairs extracted from 171,805 clinical trials. Based on these knowledge bases, we constructed two evaluation sets as the proxy of novel PD drugs: the first set consists of the drugs that have been tested for PD in clinical trials and the second set consists of PD drugs extracted from literature abstracts in Medline. We removed the FDA-approved PD drugs from both sets. We used histogram to investigate the distribution of drugs in each set among our rank. We also generated a precision-recall curve and calculated the mean average precision to evaluate the ranking of drugs in the union of the two sets.

We compared the performance of our approach with a recent drug discovery approach proposed by Hoehndorf.49 In their approach, the human diseases were linked to mouse phenotypes through phenotype ontology comparison, and then associated with orthologous genes based on the gene-phenotype relationships in animal models. After that, they linked the predicted disease genes with the drug-target data to suggest candidate drugs for a given disease. We compared the histograms that represent the distributions of evaluation drugs as well as the precision-recall curves for the two methods.

E. Test the top-ranked drugs using gene expression data analysis

We further examined the top-ranked drugs by comparing their gene expression profiles in Gene Expression Omnibus (GEO)51 with that of PD. For the drugs, we extracted data sets that contain gene expression levels before and after adding the drugs to human or animal brain tissues. For PD, we downloaded the data sets that compared the
PD patients and healthy controls. We used the GEO2R software to identify the significantly differential expressed genes (adjusted $p$ value $< 0.05$) for the disease and drugs, respectively. Then we investigated if common significant genes exist between PD and the drug, and if these common genes have opposite directions of regulation.

Results

A. Our disease genetics-based phenotype prioritization algorithm identified PD-specific mouse model phenotypes

We ranked and classified the mouse model phenotypes detected using PD genes in OMIM. The top ranked phenotype categories are nervous system and behavior/neurological phenotypes as expected (Table 1). Examples of nervous system phenotypes with the highest weights include neurodegeneration and alpha-synuclein inclusion body, which characterize the pathology of PD. In addition, top-ranked behavior/neurological phenotypes, such as impaired coordination and abnormal gait, mostly include typical motor symptoms of PD. Interestingly, the rest top-ranked phenotype categories show that the pathology of PD is complex and involves not only the nervous system, but also immune system, homeostasis and other aspects.

Table 1. The top-ranked categories of mouse phenotypes extracted using PD genes in OMIM.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Phenotype Category</th>
<th>Example top-ranked phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nervous system phenotype</td>
<td>Neurodegeneration</td>
</tr>
<tr>
<td>2</td>
<td>behavior/neurological phenotype</td>
<td>impaired coordination</td>
</tr>
<tr>
<td>3</td>
<td>immune system phenotype</td>
<td>decreased double-positive t cell number</td>
</tr>
<tr>
<td>4</td>
<td>homeostasis/metabolism phenotype</td>
<td>decreased dopamine level</td>
</tr>
<tr>
<td>5</td>
<td>hematopoietic system phenotype</td>
<td>decreased hemoglobin content</td>
</tr>
</tbody>
</table>

B. Our approach prioritized FDA-approved PD drugs

We extracted 22 FDA-approved drugs for PD and 474 genes targeted by these drugs. The median rank of the 22 drugs is 125 (top 10% among 1197 drugs). The histogram in figure 4 shows that our approach prioritized 10 approved PD drugs within top 10%. The table in figure 4 shows the rank and percentile of the top 10 approved PD drugs. Among them, the most effective dopamine replacement agent, levodopa, was ranked within top 5%. Figure 5 shows that the drugs prioritized by our approach frequently target on the drug target genes for approved PD drugs. In figure 5(a), nine in the top ten drug target genes (except GABRA1) are target genes for approved PD drugs. Figure 5(b) shows that half of the top 10% genes have been targeted by approved PD drugs, while the other half are new drug targets and may lead to novel candidate PD drugs.

Figure 4. Our approach ranked the approved PD drugs in the top. A total of 10 among 22 approved PD drugs were ranked within top 10% among all the 1197 drugs.

<table>
<thead>
<tr>
<th>Drug name</th>
<th>Rank</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>bromocriptine</td>
<td>28</td>
<td>2.34%</td>
</tr>
<tr>
<td>pergolide</td>
<td>32</td>
<td>2.67%</td>
</tr>
<tr>
<td>apomorphine</td>
<td>38</td>
<td>3.17%</td>
</tr>
<tr>
<td>ropinirole</td>
<td>39</td>
<td>3.26%</td>
</tr>
<tr>
<td>pramipexole</td>
<td>41</td>
<td>3.43%</td>
</tr>
<tr>
<td>rotigotine</td>
<td>48</td>
<td>4.01%</td>
</tr>
<tr>
<td>levodopa</td>
<td>53</td>
<td>4.43%</td>
</tr>
<tr>
<td>creatine</td>
<td>93</td>
<td>7.77%</td>
</tr>
<tr>
<td>scopalamine</td>
<td>109</td>
<td>9.11%</td>
</tr>
<tr>
<td>trihexyphenidyl</td>
<td>114</td>
<td>9.52%</td>
</tr>
</tbody>
</table>
Figure 5. The drug target genes that are most frequently targeted by our top 10% drugs. (a) The top 10 drug target genes for our prioritized drugs. (b) The distribution of target genes for approved PD drugs among all the drug target genes.

Approved PD drugs and their target genes cannot be easily detected through matching disease genes and drug target genes. We compared the performance in identifying approved PD drugs with two genetics-based drug discovery methods. Using the first method, none of the 15 PD genes directly matches the target genes for approved PD drugs and we detected zero approved drug. Using the second method, we detected one approved PD drug, rasagiline, through its target gene BCL2, which interacts with the PD gene PARK2. Though the disease genes for PD and their interacting genes do not directly provide information on the drug target genes, our approach prioritized 10 out of 22 approved PD drugs by exploiting the gene-phenotype associations in mouse models.

C. Our approach outperformed an existing approach in prioritizing novel PD drugs

The top ranked drugs generated by our approach are enriched for the novel PD drugs in the two evaluation sets (figure 6). We extracted 81 drugs from clinical trials to construct the first set, and the candidate drugs in our approach contain 69 of them. Our approach ranked a total of 22 drugs in the top 10% (3.2-fold enrichment comparing to the random cases, \( p < e^{-4} \)), and this number is 450% higher than 4 drugs in the bottom 10%. Most testing drugs (68%) in the clinical trial set were ranked within top 30%. The evaluation set based on Medline contains 102 drugs, and our candidate drugs included 85 among them. We ranked 26 within top 10% (3.0-fold enrichment comparing to the random cases, \( p < e^{-4} \)), which is a 760% increase comparing with 3 drugs in the bottom 10%. In contrast, figure 7 shows that the evaluation drugs spread out in different rank ranges when using the existing drug discovery approach based on mouse model phenotypes. Comparison between figure 6 and 7 show that our approach performed better than Hoehndorf’s approach in ranking novel PD drugs in the two evaluation sets.

Figure 6. The distribution of our ranks for two sets of novel PD drugs extracted from clinical trials and Medline texts.
Figure 7. The distribution of evaluation sets based on clinical trials and Medline texts among the ranks generated by the baseline approach based on mouse phenotypes.

The precision-recall curves in figure 8 further shows that our performance is significantly better than the previous approach. The mean average precision for our approach is 0.24, which is significantly higher than 0.16 for the Hoehndorf’s approach (p<e-11). The result means that our approach achieved higher precision averagely at all recall levels, and mostly ranked the novel PD drugs higher than the previous approach.

Figure 8. Precision-recall curves in ranking the novel PD drugs for our approach and Hoehndorf’s approach based on PhenomeNet.

D. Gene expression analysis suggests quetiapine as a potential PD drug

Among the top 10 candidate drugs, we found a set of gene expression samples available for quetiapine in GEO. We identified 61 significant genes for quetiapine from GEO series GSE45229 and 1650 significant genes for PD from GSE8397. Table 2 lists the common significantly differential genes between PD and quetiapine, as well as the direction of regulation for each gene and the logarithm of fold change. Among these genes, MAOA regulates the metabolism of neurotransmitters such as dopamine and is closely associated with PD. In addition, MAOA is not a drug target gene for quetiapine based on the drug-target data in DrugBank. The gene expression analysis suggests that quetiapine, one of the top ranked drugs, has the potential to treat PD.

Table 2. Common significantly differential genes for PD and quetiapine as well as their directions of regulation and fold change.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Quetiapine regulation</th>
<th>Log(FC)</th>
<th>PD regulation</th>
<th>Log(FC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSPB1</td>
<td>Up</td>
<td>1.5</td>
<td>Down</td>
<td>-1.4</td>
</tr>
<tr>
<td>CHORDC1</td>
<td>Up</td>
<td>0.6</td>
<td>Down</td>
<td>-1.1</td>
</tr>
<tr>
<td>MAOA</td>
<td>Down</td>
<td>-0.6</td>
<td>Up</td>
<td>0.8</td>
</tr>
<tr>
<td>MRPL15</td>
<td>Down</td>
<td>-0.6</td>
<td>Up</td>
<td>0.8</td>
</tr>
</tbody>
</table>
**Discussion**

Our study has several limitations that may be addressed in the future. Currently, we used the disease genetics knowledge in OMIM as the seeds to detect PD mouse phenotypes. We have demonstrated in several recent works that disease genes predicted by analyzing human disease phenotype networks and genetic functional relationship networks also have the translational potential in drug discovery. In the future, we will develop approaches to integrate disease associated genes in OMIM, GWAS and prediction results from computational approaches in the drug repositioning approach.

In addition, other information including human disease phenotypes, disease similarities and drug similarities may also contribute in further prioritize strong candidate drugs. In the future, we will develop approaches to seamlessly integrate these data, and investigate the candidate drugs from different angles.

Finally, our experiments tested the ranks of drugs in three evaluation sets: approved PD drugs, drugs that are tested for PD in clinical trials, and PD drug treatments extracted from biomedical literature. These evaluations tested the performance of our approach in retrieving the most successful PD drugs, and approximated its ability of detecting novel PD drugs. In the future, biomedical experiments and clinical studies are required to further demonstrate the effectiveness of the candidate drugs in treating PD.

**Conclusions**

In this study, we developed a novel drug repositioning approach to predict new drugs for Parkinson’s disease using both disease genetics knowledge and mouse model phenotypes. Our approach can identify FDA-approved PD drugs and prioritize novel PD drugs. A comparison with pure genetics-based drug repositioning approaches shows the importance of using mouse model phenotypes in identifying PD drugs. In addition, our approach outperformed a recently proposed mouse phenotype based drug discovery method through combining disease genetics with mouse model phenotypes using a novel computational approach. Further gene expression analysis on top-ranked candidate drugs suggested quetiapine as a potential PD therapy.

**Acknowledgement**

RX conceived the study. YC performed algorithm design, data analysis, and evaluation, and wrote the manuscript. All authors have participated in the study discussion and final manuscript review.

**Funding**

The project was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under the NIH Director’s New Innovator Award number DP2HD084068.

**References**


Automated Classification of Consumer Health Information Needs in Patient Portal Messages

Robert M. Cronin, M.D., Daniel Fabbri, Ph.D., Joshua C. Denny, M.D., M.S., and Gretchen Purcell Jackson, M.D., Ph.D.
Vanderbilt University Medical Center, Nashville, Tennessee

Abstract

Patients have diverse health information needs, and secure messaging through patient portals is an emerging means by which such needs are expressed and met. As patient portal adoption increases, growing volumes of secure messages may burden healthcare providers. Automated classification could expedite portal message triage and answering. We created four automated classifiers based on word content and natural language processing techniques to identify health information needs in 1000 patient-generated portal messages. Logistic regression and random forest classifiers detected single information needs well, with area under the curves of 0.804-0.914. A logistic regression classifier accurately found the set of needs within a message, with a Jaccard index of 0.859 (95% Confidence Interval: (0.847, 0.871)). Automated classification of consumer health information needs expressed in patient portal messages is feasible and may allow direct linking to relevant resources or creation of institutional resources for commonly expressed needs.

Introduction

Patients have health information needs about a variety of topics including symptom management, medication side effects, prognosis, coping, where and from whom to get treatment, and financial assistance1-10. In 2000, Jones described the multitude of ways patients would attempt to answer their questions with informatics tools during the next decade11, such as consumer sites on the world wide web and electronic mail messages between patients and physicians as well as among patients with similar conditions. These predictions have largely proven true1, 12-16. Recently, newer tools, such as patient portals, have emerged as another means of addressing consumer health information needs. Patient portals are web-based applications that enable patients to interact with their health information, healthcare systems, and providers17-19. Secure patient-provider messaging is one of the most popular functions of patient portals20-23. Several studies exploring the types of communication that occur through portal messaging have demonstrated the expression of important health and information needs involving prescription refills, interpretation of laboratory values, and requests for appointments20, 21, 24, 25. Other research has shown that clinical care can be delivered through portal messages24, 25. Patients may report new health problems, and these messages may result in further evaluation or treatment26.

Previous studies of the content of portal messages are limited in several ways. First, prior studies have analyzed only small numbers of messages almost exclusively in the primary care setting. Patient portals are now being widely deployed across specialties in many healthcare institutions27, 28. Second, most prior work has only described the communications with a narrow range of categories such as tests, appointments, symptoms, referrals, and general medical questions. Finally, previous research has employed manual analysis to characterize these messages. With millions of secure messages exchanged between patients and providers each year, better techniques are needed to understand fully the information needs expressed and health care delivered through patient portals.

Automated detection of concepts in consumer generated documents has been studied using natural language processing (NLP) techniques with standardized terminologies. Vocabularies have been developed to classify clinical and consumer generated documents including the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and the Consumer Health Vocabulary (CHV) vocabularies in the Unified Medical Language System (UMLS)29, 30. These vocabularies have been used in studies analyzing clinical text as well as patient-generated text on social media31-34. Previous work has shown that adding semantic types (STYs) to clinical questions leads to improved classification35. NLP techniques using standardized vocabularies and semantic types have not been studied in automated classification of patient portal messaging.

This prior research suggests that identification of consumer health information needs in portal messages should be possible, but to date, these techniques have not been applied to analyze portal communications. In this paper, we
describe the development and evaluation of an automated classifier that employs NLP techniques and machine learning algorithms on patient-generated secure messaging text to identify types of consumer health information needs within a patient portal.

Methods

Setting

This study was conducted at Vanderbilt University Medical Center (VUMC), a private, non-profit institution that provides primary and regional referral care to over 500,000 patients annually. VUMC is located in middle Tennessee and serves both adults and children with over 900 inpatient beds, 50,000 inpatient admissions, and over 1 million outpatient visits per year. VUMC launched a patient portal called My Health at Vanderbilt (MHAV) in 2005 and completed deployment of the portal throughout the clinical enterprise by 2008. MHAV patient portal is available to all patients who receive medical care at VUMC. MHAV provides a suite of common patient portal functions including access to selected portions of the electronic medical record, appointment scheduling, secure messaging, bill management, and delivery of personalized health information46, 37. MHAV currently has over 293,000 registered users, including more than 19,000 pediatric accounts, with over 255,000 logins per month.

Portal Messages

We collected the content of all patient-generated secure messages sent through the MHAV portal from the launch of the portal in 2005 through 2014. De-identified messages were extracted from the VUMC Synthetic Derivative (SD), a database containing a de-identified copy of all hospital medical records created for research purposes. Over 2.5 million patient-generated messages were present in the SD. One thousand individual messages were randomly selected from an equal distribution over the 10-year period of this data set for analysis in this study.

Taxonomy

Our research team developed a taxonomy of consumer health information needs and communications shown in Figure 1. This taxonomy provides a comprehensive model of the semantic types of consumer health communications. Previous literature describes classification of clinical and other healthcare provider needs38-41. Classifying consumer health information needs has been an ongoing research question1, 2, 4-10, 42-44, and some studies have examined patient and caregiver needs in selected diseases45, 46. However, existing taxonomies have been incomplete or difficult to use.

Our taxonomy divides information needs and communications into five main categories: clinical information, medical, logistical, social, and other. We use the taxonomy to describe both needs and communications because it can be employed to categorize both questions and the answers to these questions. This taxonomy has evolved from a model of clinical information needs that express questions that require medical knowledge, such as those which could be answered by a consumer health information resource40. This component of the model is the most well developed as it has been employed to structure medical textbooks and evaluated on a diverse set of communications including patient journals and questions from patient and caregiver interviews40, 47. The model was then expanded to add the medical, logistical, and social needs that are expressed in other types of communications, such as portal messages. Medical needs are requests for delivery of medical care, such as the expression of a new symptom requiring management or an inquiry about a test result. Logistical needs are requests for pragmatic information, such as the location of a clinic or the copy of a medical record. The social category includes personal communications such as an expression of gratitude or a complaint. The other category covers communications that are incomplete or unintelligible. Portal messages can contain more than one type of communication. Components of the taxonomy have been validated with inter-rater reliability studies of classification of consumer questions.47

Gold Standard

A gold standard was developed by manual analysis of the types of needs and communications that were present in the selected 1000 messages. Two to three individuals reviewed the content of all 1000 messages and assigned all relevant categories to each message. Discrepancies were discussed and consensus achieved to produce this gold standard.

Our 1000 patient generated messages contained 721 medical needs, 234 social needs, 121 clinical information needs, and 222 logistical needs (Table 1). Thirty-one of the patient-generated messages were categorized as other, which consisted of error messages, incomplete messages, or messages that were incomprehensible. The number of
different major categories in each of the remaining messages included 676 with one major category, 260 with two major categories, 30 messages with three major categories, and three messages with all four major categories. We used a co-occurrence matrix (Table 1) to represent how often information needs from multiple categories are expressed in a single message (the value in each cell represents the number of messages that have category x and y over those messages that have category x).

**Figure 1.** The taxonomy of consumer health information needs.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G. Follow up</td>
<td>13. Adverse effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**III. Logistical Needs or Communications**

| A. Appointments/scheduling |
| B. Medical equipment |
| C. Personnel/referrals |

**IV. Social Needs or Communications**

| A. Acknowledgment |
| B. Complaints |
| C. Relationship communications |
| D. Miscellaneous |

**V. Other**

**Table 1: Distribution of categories of messages**

| Distribution of categories of messages N (% of total messages) |
| --- | --- | --- |
| Need Category | Category Present | Category Absent |
| Clinical Information | 121 (12.1%) | 879 (87.9%) |
| Medical | 721 (72.1%) | 279 (27.9%) |
| Logistical | 222 (22.2%) | 778 (77.8%) |
| Social | 234 (23.4%) | 766 (76.6%) |

* The full distribution of messages for each subcategory of the taxonomy can be obtained from the authors.
Table 2: Co-occurrence matrix with the percentage of messages that occur between two categories by row (e.g., 79% of messages with clinical information need have a medical need).

<table>
<thead>
<tr>
<th>Need Category</th>
<th>Clinical Information</th>
<th>Medical</th>
<th>Logistical</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Information</td>
<td>100%</td>
<td>79%</td>
<td>12%</td>
<td>7%</td>
</tr>
<tr>
<td>Medical</td>
<td>13%</td>
<td>100%</td>
<td>19%</td>
<td>10%</td>
</tr>
<tr>
<td>Logistical</td>
<td>7%</td>
<td>63%</td>
<td>100%</td>
<td>15%</td>
</tr>
<tr>
<td>Social</td>
<td>4%</td>
<td>31%</td>
<td>7%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Automated Classifiers

Automated classifiers utilized message contents to learn the major categories of consumer health information needs present using the taxonomy described above. We built four classifiers to identify consumer health information need categories in portal messages: basic, Naïve Bayes, logistic regression, and random forest. The basic classifier used regular expressions to detect whether a few words for each consumer health need category appeared in the messages (Table 2). The three other classifiers used the following machine learning techniques to predict if health information need categories were present in the message: Naïve Bayes, logistic regression, and random forests. To create the classifiers, we used python’s Scikit Learn package. We used Bernoulli Naïve Bayes with an alpha of 0.1 and random forests with 500 trees. The inputs to the machine learning classifiers consisted of Bag of Words (BoW), concept unique identifiers (CUIs), and semantic types (STYs). BoW is a representation of messages as a bag (or set) of its words represented as a vector for each message representing the number of times a word appears in the message. CUIs are unique concepts or meanings of words and STYs are broad categories of concepts represented in the Unified Medical Language System (UMLS). All features were represented as matrices with the messages representing the rows, the different features representing the columns. For the BoW, the numbers of occurrences of each word in each message make up the cells in a row. CUIs and STYs were binary features, which were zero or one depending on whether the CUI or STY was present in the message. Common stop words were removed from messages for the BoW representation. To determine CUIs and STYs, we used Knowledge Map Concept Indexer (KMCI), a validated tool designed at Vanderbilt, to pull concepts from the text within these messages using NLP and UMLS. The machine learning classifiers were trained and tested on a gold standard corpus of 1000 documents with 5-fold cross validation.

Table 3. Basic classifier using words to determine if a message belongs to one of the major categories of health information needs.

<table>
<thead>
<tr>
<th>Need Category</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Information</td>
<td>question, normal, medication, procedure</td>
</tr>
<tr>
<td>Logistical</td>
<td>insurance, record, bill, cover</td>
</tr>
<tr>
<td>Social</td>
<td>thank you very much, thank you so much, thanks very much, thanks so much, appreciate, your time</td>
</tr>
<tr>
<td>Medical</td>
<td>refill, prescription, appointment, pain, hurt, lab, follow up, test, xray, ct, mri</td>
</tr>
</tbody>
</table>

Evaluation and Statistical Analysis

We wanted to determine the ability for classifiers to predict a single category of need in a message and to predict all of the major categories of needs within a message. We evaluated the ability to predict a single major category in the machine learning classifiers with area under the receiver operator curves (AUCs). AUCs can measure the predictive...
ability of each classifier to learn whether a single category is present in a given patient generated message. For our basic classifier, we determined the ability to identify an information need category through the presence of several typical words found in those categories within the message.

We used the Jaccard index\(^{51}\) to predict how well the classifiers are able to learn the set of information needs that are in a single message, since multiple categories of information needs may be present in a single message. The Jaccard index is a measure of the similarity between two sets:

\[
J(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]

The Jaccard Index was chosen for its ability to determine similarities between two sets of binary outcomes. It has similar performance in text classification tasks as other similarity metrics such as Pearson’s correlation coefficient\(^{52}\). A Jaccard Index of 1 indicates that the sets A and B have all elements in both sets, and a Jaccard Index of 0 means the sets A and B have no common elements. In our study, the gold standard annotated set represents A and the predicted set from the different classifiers represents B. We averaged the Jaccard indices for each message to give an overall estimation of the ability to predict the set of information needs across the entire corpus of messages. This study was approved as non-human subjects research by the VUMC Institutional Review Board.

### Results

The AUCs for the classifiers had different values depending on what type of input was used (Tables 4-7). The basic classifier’s AUCs ranged from 0.674 to 0.848. AUCs for Naïve Bayes ranged from 0.557 to 0.796, with medical needs having the highest AUC when using BoW and STYs (AUC: 0.796; 95% Confidence Interval (CI): (0.779, 0.813)). Logistic regression’s AUCs ranged from 0.814 to 0.883, with logistical needs using BoW, CUIs, and STYs having the highest AUC (AUC: 0.883; 95% CI: (0.863, 0.903)). Random forest’s AUCs ranged from 0.804 to 0.914 with logistical communication using BoW, CUIs and STYs having the highest AUC (AUC: 0.914; 95% CI: (0.883, 0.945)). The highest AUC for clinical information needs was the basic classifier (AUC: 0.848; 95% CI: (0.843, 0.853)); logistical needs was random forest with BoW and STYs (AUC: 0.914; 95% CI: (0.883, 0.945)); social needs was random forest with BoW, CUIs, and STYs (AUC: 0.839; 95% CI: 0.822, 0.855)); and medical needs was logistic regression with BoW and CUIs (AUC: 0.870; 95% CI: (0.842, 0.897)).

The Jaccard index for the basic classifier averaged over all 1000 documents was 0.674 (95% CI: (0.663, 0.684)). The average Jaccard indices for the different machine learning classifiers also were different depending on the inputs (Table 7). Logistic regression’s highest Jaccard index was 0.859 (95% CI: (0.847, 0.871)) when using BoW and CUIs with or without STYs, while random forest’s highest Jaccard index was for BoW and STYs: 0.858 (95% CI: (0.847, 0.870)). Naïve Bayes’ highest Jaccard index was 0.776 (95% CI: (0.763, 0.790) with BoW and STYs.

### Table 4

The area under the curves (AUCs) of the machine learning classifiers for *Clinical Information Needs* with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th># of Features</th>
<th>Basic Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Words</em></td>
<td>4</td>
<td>0.848 (0.843,0.853)</td>
</tr>
<tr>
<td><em>BoW</em></td>
<td>3,194</td>
<td>0.743 (0.716,0.771)</td>
</tr>
<tr>
<td><em>CUI</em></td>
<td>2,059</td>
<td>0.698 (0.638,0.757)</td>
</tr>
<tr>
<td><em>STY</em></td>
<td>141</td>
<td>0.602 (0.569,0.635)</td>
</tr>
<tr>
<td><em>BoW, CUI</em></td>
<td>5,253</td>
<td>0.754 (0.715,0.794)</td>
</tr>
<tr>
<td><em>BoW, STY</em></td>
<td>3,335</td>
<td>0.743 (0.719,0.767)</td>
</tr>
<tr>
<td><em>CUI, STY</em></td>
<td>2,200</td>
<td>0.696 (0.635,0.757)</td>
</tr>
<tr>
<td><em>BoW, CUI, STY</em></td>
<td>5,394</td>
<td>0.751 (0.711,0.790)</td>
</tr>
</tbody>
</table>
Table 5. The area under the curves (AUCs) of the machine learning classifiers for Logistical Needs with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th># of Features</th>
<th>Basic Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>4</td>
<td>0.819 (0.812,0.826)</td>
</tr>
<tr>
<td>BoW</td>
<td>3,194</td>
<td>Naïve Bayes 0.755 (0.734,0.775)</td>
</tr>
<tr>
<td>CUI</td>
<td>2,059</td>
<td>Logistic Regression 0.878 (0.841,0.915)</td>
</tr>
<tr>
<td>STY</td>
<td>141</td>
<td>Random Forest 0.900 (0.862,0.936)</td>
</tr>
<tr>
<td>BoW, CUI</td>
<td>5,253</td>
<td>Naïve Bayes 0.776 (0.755,0.796)</td>
</tr>
<tr>
<td>BoW, STY</td>
<td>3,335</td>
<td>Logistic Regression 0.876 (0.842,0.910)</td>
</tr>
<tr>
<td>CUI, STY</td>
<td>2,200</td>
<td>Random Forest 0.909 (0.879,0.938)</td>
</tr>
<tr>
<td>BoW, CUI, STY</td>
<td>5,394</td>
<td>Naïve Bayes 0.782 (0.758,0.805)</td>
</tr>
</tbody>
</table>

Table 6. The area under the curves (AUCs) of the machine learning classifiers for Social Needs with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th># of Features</th>
<th>Basic Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>6</td>
<td>0.759 (0.736,0.782)</td>
</tr>
<tr>
<td>BoW</td>
<td>3,194</td>
<td>Naïve Bayes 0.673 (0.653,0.692)</td>
</tr>
<tr>
<td>CUI</td>
<td>2,059</td>
<td>Logistic Regression 0.791 (0.756,0.826)</td>
</tr>
<tr>
<td>STY</td>
<td>141</td>
<td>Random Forest 0.810 (0.782,0.837)</td>
</tr>
<tr>
<td>BoW, CUI</td>
<td>5,253</td>
<td>Naïve Bayes 0.658 (0.629,0.687)</td>
</tr>
<tr>
<td>BoW, STY</td>
<td>3,335</td>
<td>Logistic Regression 0.795 (0.762,0.828)</td>
</tr>
<tr>
<td>CUI, STY</td>
<td>2,200</td>
<td>Random Forest 0.821 (0.789,0.854)</td>
</tr>
<tr>
<td>BoW, CUI, STY</td>
<td>5,394</td>
<td>Naïve Bayes 0.674 (0.652,0.695)</td>
</tr>
</tbody>
</table>

Table 7. The area under the curves (AUCs) of the machine learning classifiers for Medical Needs with each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest AUC is bolded for each classifier.

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th># of Features</th>
<th>Basic Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>11</td>
<td>0.674 (0.663,0.684)</td>
</tr>
<tr>
<td>BoW</td>
<td>3,194</td>
<td>Naïve Bayes 0.780 (0.765,0.796)</td>
</tr>
<tr>
<td>CUI</td>
<td>2,059</td>
<td>Logistic Regression 0.861 (0.829,0.894)</td>
</tr>
<tr>
<td>STY</td>
<td>141</td>
<td>Random Forest 0.842 (0.808,0.875)</td>
</tr>
<tr>
<td>BoW, CUI</td>
<td>5,253</td>
<td>Naïve Bayes 0.776 (0.751,0.800)</td>
</tr>
<tr>
<td>BoW, STY</td>
<td>3,335</td>
<td>Logistic Regression 0.870 (0.842,0.897)</td>
</tr>
<tr>
<td>CUI, STY</td>
<td>2,200</td>
<td>Random Forest 0.843 (0.810,0.875)</td>
</tr>
<tr>
<td>BoW, CUI, STY</td>
<td>5,394</td>
<td>Naïve Bayes 0.781 (0.757,0.805)</td>
</tr>
</tbody>
</table>

1866
Table 8. The average Jaccard indices with a 95% confidence interval of the machine learning classifiers for each type of input: Bag of Words (BoW), unique concept identifiers (CUIs), and semantic types (STYs). The highest Jaccard index is bolded for each classifier.

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th># of Features</th>
<th>Basic Classifier</th>
<th>Naïve Bayes</th>
<th>Logistic Regression</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>24</td>
<td>Basic Classifier</td>
<td>0.674 (0.663, 0.684)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BoW</td>
<td>3,194</td>
<td>Naïve Bayes</td>
<td>0.769 (0.756, 0.783)</td>
<td>0.854 (0.841,0.866)</td>
<td>0.848 (0.836,0.860)</td>
</tr>
<tr>
<td>CUI</td>
<td>2,059</td>
<td>Logistic Regression</td>
<td>0.741 (0.728,0.754)</td>
<td>0.819 (0.808,0.831)</td>
<td>0.824 (0.813,0.835)</td>
</tr>
<tr>
<td>STY</td>
<td>141</td>
<td>Random Forest</td>
<td>0.774 (0.761,0.788)</td>
<td>0.804 (0.793,0.816)</td>
<td>0.810 (0.797,0.824)</td>
</tr>
<tr>
<td>BoW, CUI</td>
<td>5,253</td>
<td>Naïve Bayes</td>
<td>0.761 (0.747,0.775)</td>
<td>0.859 (0.847,0.871)</td>
<td>0.855 (0.844,0.867)</td>
</tr>
<tr>
<td>BoW, STY</td>
<td>3,335</td>
<td>Logistic Regression</td>
<td>0.776 (0.763,0.790)</td>
<td>0.858 (0.846,0.870)</td>
<td>0.858 (0.847,0.870)</td>
</tr>
<tr>
<td>CUI, STY</td>
<td>2,200</td>
<td>Random Forest</td>
<td>0.757 (0.744,0.770)</td>
<td>0.820 (0.807,0.833)</td>
<td>0.826 (0.814,0.839)</td>
</tr>
<tr>
<td>BoW, CUI, STY</td>
<td>5,394</td>
<td>Naïve Bayes</td>
<td>0.769 (0.755,0.783)</td>
<td>0.859 (0.847,0.871)</td>
<td>0.856 (0.845,0.868)</td>
</tr>
</tbody>
</table>

Discussion

We examined the ability to classify automatically the content of patient generated messages from patient portals into consumer health information need categories. Our developed classifiers showed promise in identification of the types of consumer health information needs expressed in portal messages, but different types of needs were best identified by different approaches to automated classification. The best classifiers for each major category of information needs had high predictive ability and were able to determine which major categories are present in a single message. As adoption of patient portals increases, automated techniques may be needed to assist in managing growing volumes of secure messages. Automated classification of health information needs may aid in connecting patients to needed resources and in triaging portal messages. In addition, automated classifiers could support consumer health informatics research to understand the nature of communications and care delivered within patient portals. Such work could lead to better resources for commonly expressed information needs and might support compensation for care delivered online.

In this study, our gold standard had a majority of medical needs among messages. This finding supports previous literature demonstrating the delivery of care through patient portals. However, we did notice between 10-20% of messages discussed other content such as clinical information needs about conditions or interventions, logistical issues having to do with billing and navigating the health care system, and social needs such as acknowledgements of care. Therefore, a patient portal can be used for many different health information needs beyond those previously reported. Our classifiers had different levels of performance for each category of information needs. Certain categories, such as clinical information needs could be identified based on a few words with a basic classifier. The observed clinical information needs expressed in portal messages were typically questions about test results and procedures that have a characteristic phrasing, which could be easily identified by a few common terms. The other categories needed more sophisticated machine learning techniques as their expressions were more diverse. Multiple words appearing together in certain needs and not others would require more complex methods to identify needs. Finally, meanings of words (concepts) or categories that words fit under (semantic types) may be important in determining if a need is present. More complex methods like Naïve Bayes, logistic regression and random forests with NLP performed better in larger and more complex parts of the taxonomy.

Logistic regression and random forests performed similarly, and both outperformed Naïve Bayes. Logistic regression performed better in clinical information and medical needs, but poorer than random forests for the other categories. Each classifier performed better with different inputs. Social needs performed best when using BoW and STYs for Naïve Bayes, but BoW, CUIs, and STYs for logistic regression and random forest. BoW and STYs worked the best when determining logistical needs for random forests, but BoW, CUIs, and STYs were best for Naïve Bayes and logistic regression in the messages. However, the gains in AUCs when adding NLP CUIs and STYs over just BoW were modest in most cases. These findings likely represent the fact that NLP tools may not add
very much, and it is the combination of words that predict information needs categories. Several factors may influence this observation. Patient generated text is likely to have less formal biomedical content and thus may have fewer concepts than medical texts. Higher order NLP methods, such as negation, also likely have little impact on content type. Thus, current results do not necessarily support a significant need for NLP, but more research into NLP of patient generated messages is needed. Logistic regression identified all of the health information needs categories in a single message better than other methods based on the Jaccard index, but random forest had a similar Jaccard index. As each classifier performed better for different information needs categories, a hybrid of information needs classifiers might best determine which categories are in a single message.

The performance of these classifiers may be limited by several factors. First, patient generated messages are more likely to include misspellings which may adversely affect information need identification. These messages may also contain abbreviations not commonly used and different abbreviations for the same word. Second, automatic derivation of meaning from patient-generated texts using computers is an ongoing challenge. Our classifiers may not be able to understand the meaning of the text, and therefore cannot determine the category of information need. Third, we utilized a standardized vocabulary for our classifiers; however standardized vocabularies may not capture different ways of expressing concepts.

Automatic classification of health information needs in patient portals has several potentially important applications. First, it could allow triaging of patient generated messages to different members of the health care team or information resources. For example, logistical needs are more likely to be answerable by an administrative assistant, whereas medical needs may require a nurse or physician to respond. Clinical information needs can be answered by an information resource such as an educational module or trusted web application. Therefore, automatic classification might enable routing of these messages appropriately without human intervention. Second, classifiers might be used to detect levels of urgency in messages. North et al. showed that occasionally patients will send potentially life-threatening symptoms through patient portals21. Utilizing automated classifiers to detect urgent messages could prevent adverse events by prioritizing responses or alerting a provider through an alternative means of communication. Finally, these classifiers could be used to determine health information needs that are frequently expressed in select patient populations, drive appropriate resource development, and potentially automatically respond to messages with links to appropriate resources.

This study has several limitations. First, this study was conducted at a single institution with a locally developed patient portal. Although the information needs seen in these messages are common needs that have been seen in other papers about patient portal messaging20, 21, our results may be limited by the unique policies and procedures developed for MHAV. Second, this study employed a small data set. Therefore, all information needs and the full breadth of their expression may not be adequately represented. Third, this study has older data including the years 2005-2014, and some of the content of messages may have become antiquated based on secular trends. Our ongoing research projects will evaluate these methodologies on larger data sets, and explore the performance of automated classifiers across clinical specialties were vocabularies and needs may differ.

Conclusions
We have created automated classifiers that show promising results in identifying the types of consumer health information needs expressed in secure messages sent through a patient portal. Certain classifiers were able to determine different semantic categories of information needs better than others. Basic classifiers were better at identifying logistical needs while logistic regression and random forests were better for clinical information, social, and medical needs. NLP techniques can improve the ability to identify the types of information needs in patient-generated messages, but the improvements were modest. Additional research is needed to improve the performance of these classifiers to support applications such as message triage, question answering, resource development for common questions, and research on consumer health communications.

Acknowledgements
We are grateful to Shilo Anders, Ebone Ingram, and Jared Shenson for their assistance in creation of the manually annotated set of portal messages used as the gold standard for this research project. The portal message content used for the analyses was obtained from VUMC’s Synthetic Derivative, which is supported by institutional funding and
References


17. Patient portal - Wikipedia, the free encyclopedia.


32. Jiang L, Yang CC, editors. Using Co-occurrence Analysis to Expand Consumer Health Vocabularies from Social Media Data2013 2013: IEEE.
Health Literacy, Education Levels, and Patient Portal Usage During Hospitalizations

Sharon E. Davis, MStat1, Chandra Y. Osborn, PhD, MPH1, Sunil Kripalani, MD, MSc1, Kathryn M. Goggins, MPH1, Gretchen Purcell Jackson, MD, PhD1  
1Vanderbilt University Medical Center, Nashville, Tennessee

Abstract

Patient portal adoption has rapidly increased, and portal usage has been associated with patients’ sociodemographics, health literacy, and education. Research on patient portals has primarily focused on the outpatient setting. We explored whether health literacy and education were associated with portal usage in an inpatient population. Among 60,159 admissions in 2012-2013, 23.3% of patients reported limited health literacy; 50.4% reported some post-secondary education; 34.4% were registered for the portal; and 23.4% of registered patients used the portal during hospitalization. Probability of registration and inpatient portal use increased with educational attainment. Health literacy was associated with registration but not inpatient use. Among admissions with inpatient use, educational attainment was associated with viewing health record data, and health literacy was associated use of appointment and health education tools. The inpatient setting may provide an opportunity to overcome barriers to patient portal adoption and reduce disparities in use of health information technologies.

Introduction

Patient portals are online tools allowing patients and their families to interact with health care systems, health care providers, and portions of the patient’s electronic health record1,2. Over the last several decades, many health care organizations have adopted patient portals in response to consumer demands and regulatory pressures3-17. Patient portals have been shown to increase patient satisfaction, promote adherence to preventative and treatment recommendations, and improve clinical outcomes in the management of chronic diseases such as diabetes, hypertension, and depression12,14,17-22. There have been well-documented associations among health literacy, education levels, and the use of technology, including patient portals. In a large outpatient study of patients with diabetes, individuals with limited health literacy were less likely to use electronic health portals, a difference that persisted after adjustment for internet access23. Patients with a high school education or less were also less likely to use health portals24. Indeed, when asked about their preferences, patients with low health literacy and those with lower educational attainment expressed a preference for self-management support to be delivered in person or by phone, rather than via internet25. Complementary research shows related relationships, for example, between socioeconomic status or race and portal use26,27. However, not all have found a relationship between health literacy and the use of internet portals or mobile health tools28,29.

Nearly all patient portal research has been conducted in the outpatient setting in either primary care or medical specialty practices7,9,10. The inpatient setting, however, may represent an opportunity to explore barriers to patient portal adoption and to promote portal use30,31. The Vanderbilt University Medical Center (VUMC) has a well-established patient portal, My Health at Vanderbilt (MHAV), which has been deployed throughout all adult and pediatric specialties16. MHAV was designed and promoted for outpatient use, and some of the patient portal policies, such as a delay in availability of test results or a two to three business day expectation for secure message responses, might discourage portal usage in the acute hospital setting16. However, in recent years, both clinicians and the portal development team have observed an increase in the number patients using MHAV not only while remote from the health care system, but also when admitted to the hospital. In addition to a widely adopted portal, VUMC routinely collects health literacy data on adult patients during hospital admission intake. Integrating these data, we examined whether health literacy and educational attainment were associated with patient portal registration and portal usage in the inpatient setting.
Methods

Admission data

This study was conducted at VUMC, a comprehensive, academic health care facility in middle Tennessee and a major referral center for the Southeastern United States. We collected data on all admissions to Vanderbilt University Hospital (VUH) between January 1, 2012 and December 31, 2013. The VUMC Institutional Review Board approved this study.

Data on patient sociodemographics and clinical characteristics for all admissions were extracted from VUMC’s clinical data warehouse. We extracted information on patient age (date of birth), gender, race, date of admission, date of discharge, length of stay, chief complaint, and admitting service for each admission. Patients with “unknown” or “declined” race responses were coded as “unknown.” Anesthesiology, ophthalmology, psychiatry, research, dermatology, radiology, and radiation oncology, each of which had <100 admissions during the study period, were grouped into an “other” category. An emergency medicine category included patients who were admitted to inpatient services, but never physically left the emergency department prior to discharge. Admissions to pediatric services reflected adult patients admitted to pediatric providers at the adult hospital, which sometimes occurs for selected cases such as adult patients with pediatric cancers.

Patient portal usage

MHAV was launched in 2005 and has been implemented throughout the clinical enterprise since 2008. MHAV provides patients with electronic access to components of their health record, supports secure messaging with health care providers, allows for appointment management, and delivers personalized health education information based on patients’ diagnoses. All VUMC patients are eligible for a MHAV account, and MHAV users may grant account access to their caregivers and/or family members as delegate users. After implementation, MHAV was rolled out incrementally through the VUMC outpatient clinics. Physician champions encouraged provider use of MHAV, and technical support staff was available to patients, physicians, and staff as the portal was brought to individual clinics. Patients initially register for the portal online, with full access to electronic health record information being provided only after in-person user identity verification.

We extracted data on patient MHAV registration and use for all patients admitted to the hospital during the study period. Dates of MHAV registration were compared with dates of inpatient stay to determine whether patients had portal accounts during hospitalization. Patients were considered registered with a portal account during an admission if they were registered prior to discharge. From MHAV usage logs, we determined the functions accessed within the portal and user status (i.e., the patient or his/her delegate) during each admission. Admissions were classified as having inpatient portal use if any activity by the patient or his/her delegate occurred between the time of admission and the time of discharge. Portal activity during hospitalization was grouped into major functions: account management (e.g., assigning a delegate), appointment management, secure messaging with health care providers, viewing portions of the electronic health record (i.e., laboratory test results, radiology reports, clinical visit summaries, medication lists, immunization history), accessing health educational materials, and other (e.g., viewing website help).

Health literacy and educational attainment

Since 2010, nursing intake forms for adult admissions to VUH have included questions on educational attainment and health literacy. Years of educational attainment was collected as a continuous variable in the electronic health records by nurses, and was later categorized as less than high school, completed high school, some post-secondary education, and a 4-year college degree or more education. Health literacy was assessed using the Brief Health Literacy Screen (BHLS), a 3-item survey tool designed to identify patients with limited health literacy, which has been validated in research settings and as administered by nurses in our medical center. The three questions of the BHLS survey ask patients: 1) “How confident are you filling out medical forms by yourself?”; 2) “How often do you have someone help you read hospital materials?”; and 3) “How often do you have problems learning about your medical condition because of difficulty understanding written information?”. Patients responded to each question using a 5-point Likert scale, with the question regarding confidence filling out forms reverse coded. Responses were summed, and higher scores indicate higher levels of literacy. We operationalized health literacy as a continuous score and a binary indicator of adequate versus limited health literacy. Patients with scores ≤9 were classified as having limited health literacy.
Statistical analysis

We restricted our analysis dataset to admissions of adult patients (age ≥18 years) with known gender, documented educational attainment, complete BHLS responses reported during admission, and discharge date no later than March 31, 2014, which was the end of our follow-up period.

Unadjusted and adjusted logistic regression models examined the relationships between patient characteristics, including sociodemographics and health literacy, and 1) MHAV registration status, 2) inpatient use during admissions of registered patients, and 3) portal functions accessed during admissions with inpatient use. For the fully adjusted models, we included age, gender, race, length of stay, month of admission, and admitting service. Educational attainment and health literacy were both included in multivariable models. Continuous variables – BHLS score, educational attainment in years, age in years, length of stay in days, and month of admission – were fit with 4-knot restricted cubic splines. In all models, we applied the Huber-White method to correct standard errors for correlation among multiple admissions to the same patient. Our analyses were conducted in R 3.1.2.

Table 1. Sociodemographics and clinical characteristics of the study population by health literacy status.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Study Population by Health Literacy</th>
<th>Excluded Admissions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Adequate</td>
</tr>
<tr>
<td>Total</td>
<td>60,159</td>
<td>46,139</td>
</tr>
<tr>
<td>Age at admission (years) c</td>
<td>41/55/66</td>
<td>39/53/65</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>29,476</td>
<td>49.0</td>
</tr>
<tr>
<td>Male</td>
<td>30,683</td>
<td>51.0</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>545</td>
<td>0.9</td>
</tr>
<tr>
<td>African American</td>
<td>9,314</td>
<td>15.5</td>
</tr>
<tr>
<td>Native American/Alaskan</td>
<td>183</td>
<td>0.3</td>
</tr>
<tr>
<td>Caucasian</td>
<td>48,769</td>
<td>81.1</td>
</tr>
<tr>
<td>Unknown/Declined</td>
<td>1,348</td>
<td>2.2</td>
</tr>
<tr>
<td>Educational attainment category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No high school</td>
<td>2,853</td>
<td>4.7</td>
</tr>
<tr>
<td>Some high school</td>
<td>5,336</td>
<td>8.9</td>
</tr>
<tr>
<td>High school/GED</td>
<td>21,673</td>
<td>36.0</td>
</tr>
<tr>
<td>Some post-high school</td>
<td>16,221</td>
<td>27.0</td>
</tr>
<tr>
<td>College degree or more</td>
<td>14,076</td>
<td>23.4</td>
</tr>
<tr>
<td>Year of admission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>29,611</td>
<td>49.2</td>
</tr>
<tr>
<td>2013</td>
<td>30,548</td>
<td>50.8</td>
</tr>
<tr>
<td>Length of stay (days) c</td>
<td>2/3/5</td>
<td>1/3/5</td>
</tr>
<tr>
<td>Admitting service category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency Medicine</td>
<td>491</td>
<td>0.8</td>
</tr>
<tr>
<td>Medicine</td>
<td>30,934</td>
<td>51.4</td>
</tr>
<tr>
<td>Neurology</td>
<td>3,326</td>
<td>5.5</td>
</tr>
<tr>
<td>Obstetrics/Gynecology</td>
<td>1,119</td>
<td>1.9</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>102</td>
<td>0.2</td>
</tr>
<tr>
<td>Surgery</td>
<td>24,087</td>
<td>40.0</td>
</tr>
<tr>
<td>Other</td>
<td>100</td>
<td>0.2</td>
</tr>
<tr>
<td>MHAV registered user</td>
<td>20,671</td>
<td>34.4</td>
</tr>
<tr>
<td>MHAV use during admission</td>
<td>4,829</td>
<td>8.0</td>
</tr>
</tbody>
</table>

a Limited health literacy defined as score ≤9.  

b Admissions excluded due to missing education or health literacy data.  

c median and interquartile range (25th percentile / median / 75th percentile) included for continuous variables.
Results

Between January 1, 2012 and December 31, 2013, there were 84,185 admissions to VUH. We excluded 7 admissions with discharge dates after our follow-up period, 638 admissions of patients with age <18 years, and 5 admissions for which patient gender was not available. Of the 83,535 remaining admissions, the BHLS was not completed in 21,903 admissions (26.2%) and educational attainment was missing in another 1,473 admissions, restricting our analysis to 60,159 admissions and 41,176 unique patients. Whether health literacy and education data were record was associated with sociodemographics and clinical characteristics (see Table 1). Compared to patients with health literacy and education data, those excluded due to missing health literacy or education data were more likely to be younger, female, and Asian/Pacific Islander or unknown race (p<0.05). Patients without health literacy or education data also had shorter length of stays than patients with these data (p<0.05). Admissions to Obstetrics/Gynecology were least likely to have health literacy or education data available.

A description of the study population is presented in Table 1. The study population was predominantly Caucasian (81.1%), with a median age at admission of 55 years (interquartile range 41–66 years). Half of all hospitalizations were admitted by Medicine (51.4%) and the median length of stay was three days (interquartile range 2–5 days). Patients were registered for MHAV during 34.4% of admissions and used MHAV during inpatient stay in 8.0% of all admissions. Among admissions involving patient registered for MHAV, inpatient portal use occurred in 23.4%.

Across all admissions, 23.3% of patients reported limited health literacy (Table 2), with reading hospital materials being the most challenging question. Half of admissions involved patients with some post-secondary education (50.4%), while 13.6% involved patients with less than a high school education. Literacy levels and educational attainment differed by demographics (p<0.05). Correlation between continuous BHLS scores and educational attainment in completed years was 0.32.

Table 2. Health literacy and educational attainment by MHAV registration and inpatient use.

<table>
<thead>
<tr>
<th>Portal registration</th>
<th>Inpatient portal use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>Registered</td>
</tr>
<tr>
<td></td>
<td>Unregistered</td>
</tr>
<tr>
<td></td>
<td>User</td>
</tr>
<tr>
<td></td>
<td>Nonuser</td>
</tr>
<tr>
<td>N</td>
<td>60,159</td>
</tr>
<tr>
<td>Health literacy</td>
<td></td>
</tr>
<tr>
<td>BHLS score (mean, sd)</td>
<td>12.1 (3.3)</td>
</tr>
<tr>
<td>Limited literacy a (%)</td>
<td>23.3</td>
</tr>
<tr>
<td>Educational attainment</td>
<td></td>
</tr>
<tr>
<td>Completed years (mean, sd)</td>
<td>13.3 (3.0)</td>
</tr>
<tr>
<td>No high school</td>
<td>4.7</td>
</tr>
<tr>
<td>Some high school</td>
<td>8.9</td>
</tr>
<tr>
<td>High school/GED</td>
<td>36.0</td>
</tr>
<tr>
<td>Some post-high school</td>
<td>27.0</td>
</tr>
<tr>
<td>College degree or more</td>
<td>23.4</td>
</tr>
<tr>
<td></td>
<td>20,671</td>
</tr>
<tr>
<td></td>
<td>12.8 (3.0)</td>
</tr>
<tr>
<td></td>
<td>11.7 (3.4)</td>
</tr>
<tr>
<td></td>
<td>12.6 (3.1)</td>
</tr>
<tr>
<td></td>
<td>12.9 (3.0)</td>
</tr>
<tr>
<td></td>
<td>39,488</td>
</tr>
<tr>
<td></td>
<td>16.1</td>
</tr>
<tr>
<td></td>
<td>27.1</td>
</tr>
<tr>
<td></td>
<td>18.1</td>
</tr>
<tr>
<td></td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>4,829</td>
</tr>
<tr>
<td></td>
<td>14.5 (3.0)</td>
</tr>
<tr>
<td></td>
<td>14.3 (2.8)</td>
</tr>
<tr>
<td></td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>27.2</td>
</tr>
<tr>
<td></td>
<td>28.8</td>
</tr>
<tr>
<td></td>
<td>32.0</td>
</tr>
<tr>
<td></td>
<td>38.4</td>
</tr>
<tr>
<td></td>
<td>35.7</td>
</tr>
</tbody>
</table>

a Limited health literacy defined as score ≤9.

Portal registration

Health literacy and educational attainment were higher for admissions of patients registered for MHAV compared to admissions of patients not registered for MHAV (Table 2). Two-thirds (67.6%) of registered patients reported some post-secondary education, while only 41.3% of unregistered patients reported similar educational attainment. In unadjusted and adjusted models, the probability of portal registration increased across the range of educational attainment (p<0.05; Figure 1). The largest increase in the probability of registration occurred between 10 and 15 years of education (roughly between high school and college degrees).

Health literacy was limited in 16.1% and 27.1% of admissions of registered and unregistered patients, respectively. In both unadjusted and adjusted models for portal registration, patients with limited health literacy were less likely to be registered than patient with adequate health literacy (p<0.05). For BHLS scores above 10, the predicted probability of registration increased with BHLS score; a clear and significant pattern of association within lower ranges of health literacy was not observed (Figure 1).
Inpatient portal use

Among admissions of registered patients, admissions with and without inpatient use had similar levels of health literacy and educational attainment. For both registered patients using the portal during their inpatient stay and those registered patients not using the portal during their inpatient stay, two-thirds reported education beyond the high school level and approximately 5% reported never attending high school. After adjustment for sociodemographics, clinical characteristics, and health literacy, educational attainment remained significantly associated with inpatient portal use (p<0.05). For MHAV registered patients with approximately 15 or more years of education, the predicted probability of inpatient portal use increased with additional years of education, although confidence intervals became wide due to small sample sizes for the highest educational levels reported (Figure 2).
Limited health literacy was reported for 18.1% of admissions with inpatient portal use and 15.5% of admissions of registered without inpatient portal use; mean BHLS scores were not different between the two groups (12.6 for users and 12.9 for non-users). With adjustment for sociodemographics and clinical characteristics, we did not observe an association between health literacy and the inpatient portal use (Figure 2).

**Portal functions accessed during inpatient stay**

During the 4,829 admissions with inpatient portal use, patients viewed their health record information in 71.8% of admissions, used secure messaging in 57.2% of admissions, managed medical appointments in 21.0% of admissions, viewed health education materials in 20.1% of admissions, conducted portal account management activities in 4.5% of admissions, and used other functions in 8.0% of admissions. In adjusted models, educational attainment was associated with whether data from the electronic health record was viewed through the portal during admission (p<0.05). The predicted probability of viewing health record information through the portal steadily increased with years of education (Figure 3). We also observed, in adjusted models, associations between health literacy and whether appointment management tools and health education materials were accessed through the portal during admission (p<0.05). The predicted probability of using either function increased over the lower ranges of BHLS, peaking around scores of 10-11 and then declining somewhat (Figure 4).

![Figure 3](image.png)

**Figure 3.** Among admissions with inpatient portal use, predicted probability of viewing health record data through the portal by educational attainment, after adjustment for sociodemographics and clinical characteristics.

**Discussion**

This study is one of the first to examine patient portal usage during hospitalizations and the relationships between patient characteristics, including sociodemographics, health literacy, and education. In a large population of adult patients with diverse health care needs and diagnoses, we found that both health literacy and educational attainment appeared to have a significant effect on the likelihood of being registered for a patient portal. Probability of registration and inpatient portal use increased with educational attainment. As educational attainment increased, portal users were more likely to access health record information within the patient portal. While health literacy was associated with registration, it was not associated with whether patients accessed the portal during admission. Among inpatient portal users, however, health literacy was associated with use of appointment management and targeted health education tools. Similar to our results, Mayberry et al did not observe an association between health literacy and portal use. In contrast, Sarkar et al observed lower levels of portal use among patients with limited health literacy. Both studies were conducted in outpatient populations limited to patients with diabetes.
Figure 4. Among admissions with inpatient portal use, predicted probability of accessing appointment management tools and viewing health education materials through the portal by health literacy, after adjustment for sociodemographic and clinical characteristics.

There are several possible explanations for our findings. First, registration for the portal may be a formidable barrier that prohibits patients with limited levels of health literacy and educational attainment from becoming portal users. Prior research has also observed lower portal registration rates among patients with lower educational attainment and limited health literacy.23, 24 Once patients have overcome the registration barrier, education and health literacy may no longer affect portal usage. If this explanation were true, we expect to see similar trends in both the inpatient and outpatient settings. Studies of outpatient populations, however, report inconsistent findings regarding associations between portal use and literacy.23, 28 Alternatively, while admitted to the hospital, patients may receive encouragement and support in using the patient portal from hospital staff or family members, as well as contextual support such as continuous computer and internet access. This assistance may help patients overcome barriers resulting from limited health literacy or education. For example, during inpatient stay, all VUH patients may have access to the internet through computers available in patient rooms. Nurse may assist patients with use of in-room technologies, allowing patients to access the portal. If limited health literacy and education are associated with more limited computer or internet access in patients’ daily lives, then equity in computer access during an inpatient stay may have a clear connection to our results. For patients with limited access to technology in the outpatient setting, portal registration may be more difficult, or even limited to times during which patients interact with the health system, leading to associations such as those we observed between education, health literacy, and registration. During hospitalizations at VUMC, however, this technology access barrier is eliminated, which may explain the lack of an association we observed between health literacy and inpatient portal use. Whether registration serves as the prohibitive barrier to portal adoption or the presence of additional portal support during an inpatient stay is responsible for the observed associations, the inpatient setting may provide an opportunity to promote patient portals and reduce disparities in adoption of patient engagement technologies.

This study is strengthened by leveraging routinely-collected health literacy data obtained on a large cohort of adult patients; exploring adoption of a well-established patient portal with a large user community; and including patients treated by all clinical specialties rather than limiting to patients with particular diseases. However, the study is not without limitations. MHAV is a locally-developed patient portal deployed at a single institution and designed to support outpatient use cases. The findings for our portal may not necessarily translate to other portals with different registration processes and user interfaces, variations that may result in different barriers and associations between education, health literacy, and portal adoption. A number of other factors not examined in this study influence portal use, including socioeconomic status and whether or not the patient was an existing VUMC patient at admission. VUMC is a large referral center in the region, and thus patients who are transferred here might not have a portal account since they are unfamiliar with the medical center and most likely do not plan to return. Additionally, we
were not able to ascertain how patients accessed the portal. During the study period, VUMC did not provide dedicated computers for patient use in patient rooms, and thus patients wanting to access the portal during hospitalization would have needed access to a personal web-enabled device. More recently, VUMC has implemented an interactive television system that provides patients with in-room web access to the MHAV portal. Follow-up studies will examine whether this increase in technology access is associated with increased inpatient portal use. Additionally, we did not account for severity of admitting diagnoses or affected systems, which may limit patient physical and mental ability to use the patient portal during a hospital stay.

While we can distinguish portal activity conducted through patient and delegate user accounts, we do not have health literacy or educational information for delegates. For the inpatient portal use models, we associated delegate activity during each admission with the admitted patient’s education or health literacy information, which may or may not be a reflection of the delegate’s traits. We observed delegate use in 4.2% of admissions with inpatient use, and 2.9% of admissions with inpatient use solely involved delegate activity and no portal access through the patient’s account. While the delegate account policy is designed to avoid multiple users accessing the patient account, we cannot be certain whether patients used the portal themselves or logged in to allow family members access to their health information, in which case the recorded education and health literacy information may not be representative of the individual interacting with the patient portal.

Additionally, almost one fourth of patients were missing health literacy or education data. Although this limited our sample size, the remaining cohort of patients was still large. As we continue to administer health literacy measures to patients, we will be able to add health literacy and education data for patients from previous admissions. Finally, these measures are beginning to be administered in the clinics at VUH, making clinic intake forms another valuable resource of health literacy and education data for patients.

Conclusion

Patient portals engage patients in their health care, increase patient self-efficacy, and improve the quality of care. While studies of portal adoption have focused on outpatient primary care settings, patient portals may provide valuable benefits to hospitalized patients. Leveraging data from a patient portal with a large population of registered users treated across inpatient clinical specialties and data from a health literacy assessment widely implemented throughout the adult inpatient setting, we sought to understand associations between patients’ characteristics, including sociodemographics and health literacy, and portal registration and inpatient portal use. In our diverse adult inpatient population, lower education and limited health literacy were independently associated with a reduced probability of being registered for a portal account. Among patients registered for the portal, however, health literacy was not associated with whether patients accessed the portal during a hospital stay. The inpatient setting may serve as an opportunity to promote patient portal adoption and increase patient confidence with portal tools, increasing patient engagement during admission and potentially maintaining this engagement as patients transition to outpatient care. A deeper understanding of how education and health literacy act as barriers to portal registration more so than portal use is necessary for designing and implementing successful patient portal promotion activities.

Acknowledgements

We thank Travis Harper and Zhou Yan for their assistance with the collection of the admission and My Health at Vanderbilt usage data respectively. Ms. Davis was supported by supported by the 5T15LM007450-12 training grant from the National Library of Medicine. Dr. Osborn was supported by a career development award K01DK087894, and R01DK100694-01A1.

References


Using High-Fidelity Simulation and Eye Tracking to Characterize EHR Workflow Patterns among Hospital Physicians

Julie W. Doberne, BS1, Ze He, MS2, Vishnu Mohan, MD, MBI1, Jeffrey A. Gold, MD1,3, Jenna Marquard, PhD2, Michael F. Chiang, MD1,4
Departments of 1Medical Informatics & Clinical Epidemiology, 3Pulmonology & Critical Care, and 4Ophthalmology, Oregon Health & Science University, Portland, OR; 2University of Massachusetts Amherst, Amherst, MA

Abstract
Modern EHR systems are complex, and end-user behavior and training are highly variable. The need for clinicians to access key clinical data is a critical patient safety issue. This study used a mixed methods approach employing a high-fidelity EHR simulation environment, eye and screen tracking, surveys, and semi-structured interviews to characterize typical EHR usage by hospital physicians (hospitalists) as they encounter a new patient. The main findings were: 1) There were strong similarities across the groups in the information types the physicians looked at most frequently, 2) While there was no overall difference in case duration between the groups, we observed two distinct workflow types between the groups with respect to gathering information in the EHR and creating a note, and 3) A majority of the case time was devoted to note composition in both groups. This has implications for EHR interface design and raises further questions about what individual user workflows exist in the EHR.

Introduction
Electronic health records (EHRs) have become a central component of the modern clinical workflow, serving as a central documentation repository, an ordering mechanism, and a provider communication tool. EHRs have been promoted as a mechanism for improving quality of healthcare delivery, patient safety and provider efficiency. Widespread adoption has been driven in part by substantial governmental incentives through the Centers for Medicare and Medicaid Services (CMS) Meaningful Use program, with 59% of U.S. hospitals and 48% of office-based providers using EHRs as of 2014.

Implementations of EHRs have been shown to dramatically influence clinical workflows. End-user behaviors and training approaches are highly variable. Adaptive end-user behaviors such as excessive use of copy-paste/copy forward, “overdocumentation”, and “upcoding” may compromise health care quality and patient safety; training and standardization may help reduce these practices. Beginning efforts have been made to establish standard practices among EHR end users. However, these efforts have largely focused on documentation and not on information review.

The challenges facing consistency in EHR training and use are diverse. Though back end databases are fairly consistent across instances of an EHR, user interfaces and workflows can be substantially different depending upon the institution and clinical environment in which the EHR is used. Within an institution or practice group, the physician-level characteristics in usage of EHR features and usage intensity have been found to be highly variable and personalized.

The strong influence of personal experiences and preferences is thought to partly explain this variance. Assessment of end-user EHR behaviors has often been conducted via self-reports and surveys, direct observation, and meaningful use measure reporting. Survey and reporting methods can provide a high-level perspective of provider behavior, but do not capture individual workflows. Direct observation may lack the accuracy required to quantitatively evaluate differences between individuals or groups of individuals. Real-time eye tracking technology has been shown to successfully capture user behavior in online website searching, website interface design, visual attention and video games. Within the realm of medicine, it has been employed to study radiologic and electrocardiography interpretation, note reading, and medication administration. The purpose of this study is to address this gap in knowledge by characterizing the workflow patterns of physicians using the EHR. This is done using a mixed methods approach employing a high-fidelity EHR simulation environment equipped with eye and screen tracking, surveys, and semi-structured interviews to characterize the typical EHR usage by a group of hospital physicians (hospitalists) as they encounter a new patient.
Methods

This study was approved by the Institutional Review Board at Oregon Health & Science University (OHSU; Portland, OR). Subjects signed a consent form prior to participation.

Development of Simulation Cases

An instance of our EHR environment (EpicCare, Epic Systems Inc., Verona, WI) was created to house simulated patient cases. This simulation environment imports all end-user customizations from the actual EHR environment, so the interface looked exactly as it would for each subject. Simulated cases were based upon real patient cases with common principal diagnoses (i.e., among the top 10 most common ICD-9 diagnoses for adults upon hospital discharge). Two patient cases (Cases A and B) were created and independently reviewed for medical accuracy and clinical realism by domain experts in accordance with previously published recommendations for high-fidelity case creation. Both patients had previously established care at our institution, but were now presenting to the emergency department (ED) for evaluation of a new set of symptoms. Each case contained historical data in each of the categories listed in Table 1; the current ED visit contained vitals, intake/output, laboratory data, EKG, chest roentgenogram, and a half-completed ED resident note stating the history of present illness, physical exam findings, and review of systems.

Recruitment and Testing of Subjects

Attending physicians from the OHSU Division of Hospital Medicine comprised the study population. Simulations were conducted on a representative active patient ward to mimic external distractions encountered by physicians as they use an EHR. Subjects were asked to act as the admitting hospitalist, review both patient charts, and create a history and physical (H&P) note complete with assessment and plan for each patient. Simulation time was not limited. Case order was held constant for all subjects throughout the study. After completing the cases, subjects were asked to verbally describe their typical workflow for admitting a patient. Semi-structured interview questions were used to elicit details about when they use the EHR during that process, their principal sources of information, note writing strategies, and the nature of the patient interaction. Lastly, subjects were asked to complete a questionnaire regarding demographic information, EHR experience, and general computer experience.

Eye and screen tracking were conducted using a Tobii X2 60 Eye Tracker (Tobii Systems, Danderyd Municipality, Sweden), a non-invasive tracker mounted below the computer monitor. All testing was conducted using a standardized computer station with consistent and static screen and chair height. Before each simulation, the eye tracker was calibrated to each subject using a 1-minute 9-point calibration algorithm provided by the manufacturer. Upon commencement of the simulation the screen tracking software (Tobii Studio, Tobii Systems, Danderyd Municipality, Sweden) captured screen video, keystrokes, mouse clicks, ocular saccades, and eye fixations. A velocity threshold identification filter was used to identify sets of fixations (gazes), using the standard definition of a fixation as lasting a minimum of 100 ms. Each video was coded manually by a member of the research team (JD). Videos were coded by recording the information type upon which the gaze was situated at each second of the case.

Data Analysis

Simulation gaze data were divided into two major categories: informational and navigational. Informational gazes pertained to any kind of clinical data (all entries in Table 1 except 14); navigational gazes were defined as lacking clinical data, and frequently occurred on toolbars, menus, and up/down scrolling arrows. Documentation was considered a subset of the informational gaze category. Comparisons between group means were conducted using two-sided t-tests.

<table>
<thead>
<tr>
<th>1. Social history</th>
<th>12. Intake/output</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Laboratory values,</td>
<td>13. Documentation (note)</td>
</tr>
<tr>
<td>pathology, microscopy,</td>
<td></td>
</tr>
<tr>
<td>cytology</td>
<td></td>
</tr>
<tr>
<td>4. Procedure notes</td>
<td></td>
</tr>
<tr>
<td>5. Vital signs and weight</td>
<td></td>
</tr>
<tr>
<td>6. Outside records</td>
<td>15. Past surgical history</td>
</tr>
<tr>
<td>7. Other</td>
<td></td>
</tr>
<tr>
<td>8. Past medical history</td>
<td>16. Medication list</td>
</tr>
<tr>
<td>9. Imaging results and EKGs</td>
<td></td>
</tr>
<tr>
<td>10. Outpatient clinic note</td>
<td></td>
</tr>
<tr>
<td>11. Operative reports</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Twenty-one information types used in video coding
Several metrics were used to evaluate the data. First, we measured the average duration of each participant’s gaze on each information type (Table 1), calculated by their total duration of gazes on that information type divided by the number of gazes on that information type during the case. Next, we calculated the total number of transitions between information types for each patient case, and the total number of informational gazes for each case.

Equations used in DTMC analysis

\[
P_{ij} = \Pr(X_1 = s_j|X_0 = s_i) \quad (1)
\]

\[
P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}. \quad (2)
\]

Figure 1. Equations used in DTMC analysis

We used first-order Discrete Time Markov Chains (DTMCs) to model transitions between information types. A Markov Chain consists of a series of successive state-to-state transitions (Equation 1, Figure 1), which form a transition matrix. Equation 2 illustrates an example of a transition matrix, which has three states. The row i (i = 1,2,3) shows the transition probability distribution from information-type i to other information-types j. For example, \( p_{13} \) is the probability that a subject transitions from information-type 1 to 3. Higher probabilities in the matrices indicate the information-type pairs that are more likely.

Results

Subject characteristics

From a total of 23 eligible subjects, 17 (74%) subjects completed a total of 33 patient cases. Fifty-nine percent of subjects were male; the mean length of time since medical school graduation was 13.3 years. One hundred percent of subjects described themselves as “somewhat” or “very experienced” with computers. The mean length of time using the study EHR (EpicCare) was 6.5 years.

Simulation characteristics

The subjects were divided into two groups based upon how long into the case it took them to begin composing a note. This division was based upon a natural grouping observed in average note start times per subject (Figure 2). Subjects in Group 1 (n=8) began composing a note on average less than 2 minutes into the case; subjects in Group 2 (n=9) began composing a note on average more than 2 minutes into the case.

<table>
<thead>
<tr>
<th>Subject Characteristics</th>
<th>Group 1</th>
<th>Group 2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender, % male</td>
<td>87.5%</td>
<td>33.3%</td>
<td>0.02*</td>
</tr>
<tr>
<td>Self-rated computer experience, level</td>
<td>2.5</td>
<td>2</td>
<td>0.03*</td>
</tr>
<tr>
<td>Years since medical school graduation</td>
<td>15.3</td>
<td>11.5</td>
<td>0.42</td>
</tr>
<tr>
<td>EpicCare experience, years</td>
<td>7.3</td>
<td>5.7</td>
<td>0.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation Characteristics</th>
<th>Group 1</th>
<th>Group 2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transitions</td>
<td>75.5</td>
<td>57.5</td>
<td>0.04*</td>
</tr>
<tr>
<td>Gazes</td>
<td>81.7</td>
<td>62.3</td>
<td>0.04*</td>
</tr>
<tr>
<td>Documentation, number of gazes</td>
<td>32.9</td>
<td>21.4</td>
<td>&lt;0.01*</td>
</tr>
<tr>
<td>Navigation, number of gazes</td>
<td>17.6</td>
<td>17.8</td>
<td>0.47</td>
</tr>
<tr>
<td>Number of unique information types</td>
<td>12.5</td>
<td>12.4</td>
<td>0.85</td>
</tr>
<tr>
<td>Case length, mm:ss</td>
<td>25:29</td>
<td>24:29</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 2. Demographic and simulation characteristics

The proportion of men in Group 1 (87.5%) was significantly greater than the proportion of men in Group 2 (33.3%, p=0.02) (Table 2). Using a Likert-type scale to assess self-rated computer experience (1, less experienced; 2, somewhat experienced; 3, very experienced), Group 1 reported a higher mean experience score compared to Group 2 (2.5 and 2, respectively; p=0.03). Time since medical school graduation and length of EpicCare experience did not differ significantly between the groups.

The average time for Case A was significantly longer than Case B (mean±SD, 28:12±8:05 and 21:56±6:35 respectively, p=0.02). To evaluate a potential learning effect, the note start times were normalized to total case times and compared among the groups from Case A to Case B. There was a slight difference in the decrement in note start time ratio between the groups, with Group 2’s note start ratio dropping more in Case B, but this was not significant.
Number of transitions per second was evaluated as well. Whereas the transitions per second for Group 1 decreased slightly from Case A to Case B, the transitions per second for Group 2 increased slightly, but the difference between the two was not statistically significant (G1 = 0.006, G2 = -0.005; p = 0.14).

Group 1 had significantly more gazes (G1=81.7, G2=62.3; p=0.04) and transitions (G1=75.5, G2=57.5; p=0.04) over the course of each case. There was no difference between the groups in average case time, navigation time, and average number of unique information types accessed within each case.

Semi-structured interviews

Self-described workflows elicited from the semi-structured interviews were consistent with categorization into Group 1 and Group 2. Representative quotations are shown in Table 3.

Clinical content

The information types with the longest total gaze durations are shown in Figure 3. For both Groups 1 and 2, total duration of gazes for documentation (composition of the H&P note) was much higher than all non-documentation information. G1=851 seconds, G2=745 seconds. The non-documentation information types with the greatest average and total duration of gazes (greater 10 seconds) were imaging results, inpatient progress notes, lab values, medications, and ambulatory clinic notes. These five information types were also the most often visited throughout the simulations. Group 1 gazed at laboratory values significantly more often than Group 2 (G1=13.3, G2=8.2, p=0.02). Differences were observed in less frequently visited information types as well. Past medical history (G1=0.75, G2=1.6, p=0.06), problem list (G1=1.4, G2=2.9, p=0.08), family history (G1=0.7, G2=0.2, p=0.04) and other information types (G1=3.88, G2=2.53, p=0.04) all showed slight differences in visitation frequency between groups. These trends in frequency remained consistent when the gaze values were normalized by the total number of visits to all information types.

<table>
<thead>
<tr>
<th>Time Subject Started Writing Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Subjects</td>
</tr>
<tr>
<td>Group 1</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>75</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>125</td>
</tr>
<tr>
<td>150</td>
</tr>
<tr>
<td>175</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>&gt;200</td>
</tr>
</tbody>
</table>

Figure 2. Distribution of note start times

Table 3. Representative quotations from semi-structured interviews

**Group 1**

Early note composition, greater frequency of transitions between information types

- “I often start my note right away as I go about my chart review.”
- “I tend to be linear… I jump around.”
- “I usually start with a note because it autopopulates with the information I need.”

**Group 2**

Information review with longer duration per screen and less transitions, followed by note composition

- “I review all the current data, labs, and imaging. Then look at last clinic note, meds, and clinical history. I start writing a note after that.”
- “I look at the meds, prior notes and imaging, then start putting a note skeleton together.”
- “I do a quick review of the chart before I go see the patient. Then I build my problem list and when I create the note, it auto-imports the information I’ve collected.”
Transition visualizations

Figures 4a and 4b show circle visualizations of normalized Markov Chain frequencies of information types for both groups. Nodes situated around the rim of the circle represent the various information types, ordered by size moving counter-clockwise. The sizes of the nodes are proportional to their gaze number distribution frequencies, and the thickness of the lines connecting nodes indicate the normalized frequencies of transitions (transition probability) between the two information types. For clarity, only the top 80% of total transitions are depicted here.

The general patterns are the same between the two groups: most of the transitions are centered around documentation, which is also the most frequently visited information type. We can identify the top 5 most frequently visited information types, which are consistent between the two groups: documentation, lab values, inpatient progress note, vital signs and imaging results. In addition, because documentation played a larger role in group 1 than it did in group 2, there were fewer high frequency transitions compared with group 2.

One important observation from the visualization is which information types were closely related. For example, the social history, past medical history, family history, and past surgical history nodes are co-located because they were similar in visitation frequencies. In Group 1, there are more bold lines from documentation to other information types, capturing the notion that Group 1 transitioned more frequently between documentation and other information types.

Discussion

This study assessed characteristics of physicians’ information search patterns in an EHR as they created a note for a new patient. There were two discrete types of
users based upon information review and documentation tasks. The key findings from our analyses are: 1) There were strong similarities across the groups in the information types the physicians looked at most frequently; 2) While there was no overall difference in case duration between the groups, we observed two distinct workflow types between the groups with respect to gathering information in the EHR and creating a note; and 3) A majority of the case time was devoted to note composition in both groups.

Both groups showed the same preferences for a small subset of information types, in terms of how long they looked at each information type (Figure 3). Imaging results, progress notes, laboratory values, medications, and prior clinic notes were looked at longest. The total number of unique information types also did not differ between groups. This suggests some uniformity in clinical reasoning that may be explained by their mutual medical specialty, common clinical environment, and/or similarities in medical training. Differences between groups were only found in lower-frequency information types; members of Group 1 spent less time reviewing past medical history and problem lists, and more time reviewing the family history.

Groups 1 and 2 exhibited significantly different workflow types, despite no overall difference in case completion time. Group 1, characterized by early note creation, transitioned frequently between information types in the EHR after starting the note (Figure 4). Group 2 physicians, characterized by later note creation, tended to dwell on information longer before starting to compose the note. Group 1 showed a markedly higher number of transitions and gazes compared to Group 2, confirming a higher rate of switching from one information type to another. We found significant differences between the groups in how the simulation time was used. Overall, Group 1 spent substantially longer time in the documentation phase of the simulation. Subjects’ self-described workflows (Table 3) supported a dichotomy between early and note creation. Subjects in Group 1 mentioned starting a note as one of their first activities prior to information gathering (and “jumping around” in the EHR when gathering information) whereas Group 2 members described reviewing a variety of information prior to note creation.

Both groups spent a much higher amount of time on note composition than any other task, including reviewing clinical information (Figure 3). The time and burden associated with documentation is noted in the literature. What this study highlights is the finding that documentation time may overshadow all other tasks, including time to read or review the clinical data. This raises the question of whether the time and burden of documentation relates to the untoward effects of EHR implementation on patient care that have been observed. Further research is needed to elucidate the interplay between EHR documentation burden, clinical reasoning, and patient care.

This research raises several questions about the nature of EHR information seeking and patterns in end-user behavior. Several models have been proposed to describe this process. Traditionally information seeking is viewed as a sense-making process in which the user formulates a personal perspective through finding meaning. A common thread shared by most proposed models of information seeking is that it is a dynamic process, where the users move non-linearly through levels of certainty based upon information encountered and judgments of relevancy and specificity. The resultant perspective or decision is not necessarily the same among individuals and is dependent upon the effectiveness of the user’s information retrieval. This becomes problematic in the realm of clinical medicine, where standard of care dictates that there be some baseline level of uniformity in clinical reasoning to ensure patient safety. Ensuring some baseline level of competency in information retrieval becomes crucial when considering the complexity of modern EHR systems, where it has already been demonstrated that end-user behavior is highly variable. This research suggests that users take different pathways to arrive at a common endpoint. The information used and the time to complete the task may not differ, but the order in which the end product (clinical note) is created may differ depending on the user. Of Nielson’s five criteria for usability is affordance. The different workflows described in this study support the need for interfaces to afford for 1) fluctuation between varying levels of certainty in the meaning-finding process and 2) a variety of approaches in clinical documentation.

Of note, significant differences were observed in the gender composition and self-rated computer experience level between Group 1 and Group 2. With respect to the gender difference, though the sample size of the current study is too small to draw firm conclusions from this observation, our results raise the question of whether different genders may approach the tasks of EHR information gathering and documentation differently. Gender differences in clinical reasoning and information processing have been explored previously. Myers-Levy’s theory of selectivity and information-processing research in other disciplines suggest that women are more likely to employ elaborative information processing strategies regardless of the task complexity, whereas men are more likely to utilize heuristic processing strategies, only switching to elaborative strategies on more complex tasks. Conversely, in the medical literature no significant gender-related difference has been found in diagnostic reasoning. With the insertion of the EHR into clinical workflows we must consider the electronic interface as an additional layer of complexity. Research
conducted on website audiences has shown that differences in perception and satisfaction can vary greatly among gender groups\textsuperscript{51}. Females have demonstrated greater proficiency in computer display navigation and optical cue responsiveness on the screen\textsuperscript{52}. Taken together, prior research suggests that there may be several factors that contribute to gender differences in EHR usage; further research is needed. With respect to the difference in self-rated computer experience, it is unclear whether this outcome is independent of the strong gender differences between the groups. Conflicting research exists on gender differences in perceptions of self-efficacy and attitudes toward computers\textsuperscript{48,53}. If it is an independent outcome, it is not clear that the higher level of computer experience in Group 1 translated into better performance in the simulation task.

There are several important limitations to this study. First, though simulations were conducted in a high-fidelity environment, subjects were aware that the cases were fictional patients. There were no actual patients to interact with and subjects relied upon the content of the history of present illness, review of systems, and physical exam as it was documented in the EHR. Though this may have diminished the realism of the cases, it is expected that it would exert a uniform effect on the subjects. Cases were made to be slightly less complicated than the “average” patient seen at our institution, a tertiary care facility, for the purposes of simulation time. Second, interpretation of the eye tracking gaze data is still in its infancy in clinical informatics research, and it is unclear how well gazes and transitions represent clinical reasoning processes. Third, the notes created during the simulations were not evaluated for accuracy and completeness, thus we cannot comment on whether differences in search patterns affect clinical reasoning and medical decision making. Lastly, the study was conducted among one specialty of physicians at one institution.

Conclusion

This study demonstrates the presence of two information-gathering and documentation workflows among hospitalists using the EHR to admit a new patient. This has important implications for EHR interface design, specifically with respect to affordances for multiple information-gathering pathways. Future studies must continue to examine the workflow differences among individuals, specifically pertaining to note quality, clinical accuracy, and efficiency.

References


Learning a Severity Score for Sepsis: A Novel Approach based on Clinical Comparisons

Kirill Dyagilev, PhD\(^1\), Suchi Saria, PhD\(^{1,2}\)

\(^1\)Dept. of Computer Science, Johns Hopkins University, Baltimore, MD
\(^2\)Dept. of Health Policy & Mgmt., Johns Hopkins University, Baltimore, MD

Abstract
Sepsis is one of the leading causes of death in the United States. Early administration of treatment has been shown to decrease sepsis-related mortality and morbidity. Existing scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE II) and Sequential Organ Failure Assessment scores (SOFA) achieve poor sensitivity in distinguishing between the different stages of sepsis. Recently, we proposed the Disease Severity Score Learning (DSSL) framework that automatically derives a severity score from data based on clinical comparisons — pairs of disease states ordered by their severity. In this paper, we test the feasibility of using DSSL to develop a sepsis severity score. We show that the learned score significantly outperforms APACHE-II and SOFA in distinguishing between the different stages of sepsis. Additionally, the learned score is sensitive to changes in severity leading up to septic shock and post treatment administration.

Introduction
Sepsis — a whole-body inflammatory response to infection — is one of the leading causes of death in the inpatient setting and is associated with significantly higher costs of care\(^1\). The risk of sepsis-related adverse outcomes can be reduced by early treatment\(^2\). Thus, changes in a patient’s health status is regularly assessed by the caregiver (clinicians and nurses) to plan timely interventions. This requires the caregiver to interpret a diverse array of markers (e.g., heart rate, respiratory rate, blood counts, and serum measurements) that measure the underlying physiologic and metabolic state. Such clinical assessments are time-consuming, require extensive experience, and more importantly, are prone to missing signs of decline that may be only subtly visible in the observed markers.

In this paper, we address the problem of quantifying (scoring) the latent sepsis severity of a patient at a given time. That is, we derive a mapping from the high-dimensional observed marker data to a numeric score that tracks changes in sepsis severity over time — as health worsens, the score increases, and as the individual’s health improves, the score declines. More generally, a means for accurate estimation and tracking of an individual’s health status can enable clinicians to detect critical decline such as decompensations, and acute adverse events in a timely manner. Additional potential benefits include assessment for whether an individual is being responsive to therapy, stratification of patients for resource management, and risk adjustment for clinical research\(^3\).

Background: Qualitatively, the concept of disease severity score has been described as the total effect of disease on the body; the irreversible effect is referred to as damage, while the reversible component is referred to as activity\(^4\). The precise interpretation of concepts of damage and activity are typically based on the application at hand. Desirable properties of a severity scale include: 1) face and content validity i.e., the variables included are important and clinically credible, and 2) construct validity i.e., the scoring system parallels an independently ascertained severity measurement\(^4\).

Historically, severity scores have been defined and derived in a number of different ways\(^5\). One approach is to have clinical experts fully specify the score. Namely, using the existing clinical literature, a panel of experts identifies factors that are most indicative of severity of the target disease. These factors are weighted by their relative contribution to the severity and summed together to yield the total resulting score. For example, the Acute Physiology And Chronic Health conditions score\(^6\) (APACHE II), which assesses the overall health state in an inpatient setting, uses factors that are most predictive of mortality. For instance, a heart rate between 110 and 139 beats per minute adds 2 points to the final score while a heart rate higher than 180 beats per minute adds 4 points. Similarly, mean arterial blood pressure between 70 and 109 mm Hg adds no points while a value between 50 and 69 mm Hg adds 2 points. A number of additional widely used scoring systems have been designed in this way, including the Multiple Organ Dysfunction Score\(^7\) (MODS), the Sequential Organ Failure Assessment\(^8\) (SOFA), and the Medsger’s scoring system\(^4\).

A second approach commonly taken is to characterize severity as a predictive score that is associated with the risk of experiencing a target downstream event. That is, high severity states are more likely to be associated with adverse
events and higher mortality rates. To train such a score, supervised learning is used with the presence or absence of the downstream adverse event (e.g., septic shock or mortality) as the labels. For example, the pneumonia severity index (PSI) uses this approach to combine 19 factors including age, vitals and laboratory test results, to calculate severity as the probability of morbidity and mortality among patients with community acquired pneumonia. The relative weight of each factor in the resulting score was derived by training a logistic regression function that predicts risk of death in the following 30-hour window. Others have similarly used downstream adverse events such as development of c.diff infection, septic shock, morbidity, and mortality as surrogates for severity scores. Yet other approaches use probabilistic state estimation techniques.

However, the above-mentioned approaches for the derivation of disease severity scores have their limitations. The expert-based approach captures known clinical expertise well, but it does not extend to populations where current clinical knowledge is incomplete. The approach of regressing against the risk of downstream adverse event yields scores that suffer from bias due to treatment-related censoring. Specifically, scores learned from data assuming one standard of care, when used under a different standard of care may lead to erroneous conclusions about the individual disease severity state. Say, we have data from a unit where children with temperature of 102°F are frequently prescribed treatment for the flu. This intervention subsequently cures the patients and they do not experience any associated adverse events (e.g., sepsis or septic shock) or death. However, it is also known that a child with temperature of 102°F, if left untreated, is likely to die. Thus, a temperature of 102°F is clinically considered to be a high severity state which in this unit is frequently treated in a timely manner. Since the learning algorithm uses the presence of a downstream adverse event as a surrogate marker for severity, rare occurrence of adverse events following 102°F causes the learning algorithm to score the temperature of 102°F as benign. This is problematic for two reasons. First, it is not consistent with how severity is interpreted clinically. Moreover, if such a score were to be used to guide interventions, it might erroneously cause the clinician to undertreat children with temperature of 102°F, thus likely worsening outcomes. See Dyagilev and Saria for a numerical example.

In recent work, we proposed an alternative Disease Severity Score Learning (DSSL) framework for learning severity scores. The DSSL framework leverages this key observation that, while requesting experts to quantify disease severity at a given time is challenging, acquiring clinical comparisons – clinical assessments that order the disease severity at two different times – is often easy. From these clinical comparisons, DSSL learns a function that maps the patients observed feature vectors to a scalar severity score. With some abuse of terminology, we refer to this mapping function as the DSS function. We showed empirically that the generalization performance of DSSL is less sensitive to variations in treatment administration patterns compared to the supervised learning approach described above.

In this paper, we test the feasibility of using DSSL to develop a sepsis severity score. We use the MIMIC-II dataset containing electronic health record data from patients admitted to the Beth Israel Deaconess Intensive Care Units between 2001 and 2008. Clinical comparisons required for training the score are generated automatically using the Surviving Sepsis Campaign (SSC) guidelines. We show that the learned score significantly outperforms APACHE-II and SOFA in distinguishing between the different stages of sepsis. Additionally, the learned score is sensitive to changes in severity leading up to septic shock and post treatment administration. We also show that only a small number of clinical comparisons are needed to obtain a high quality score. Thus, by not having to rely on needing a large number of clinical comparisons, the use of DSSL in domains where guidelines are unavailable for generating clinical comparisons in an automated manner (as we do in this paper) becomes feasible.

Methods
In this section, we describe the DSSL framework and the L-DSS method for learning linear DSS functions.

General DSSL Framework We consider longitudinal data routinely collected in a hospital setting. These include co-variates obtained at the time of admission such as age, gender, and clinical history; time-varying measurements such as heart rate and respiratory rate; and text notes summarizing the patients evolving health status. These data are processed and transformed into tuples \( < x_i^p, t_i^p > \) where \( x_i^p \in \mathbb{R}^d \) is a \( d \)-dimensional feature vector associated with patient \( p \in P \) at time \( t_i^p \) for \( i \in \{1, \ldots, T_p\} \) and \( T_p \) is the total number of tuples for patient \( p \). A feature vector \( x_i^p \) contains raw measurements (e.g., last measured heart rate or last measured white blood cell count) and features derived from one or more measurements (e.g., the mean and variance of the measured heart rate over the last six hours or the total urine output in the last six hours per kilogram of weight). Let \( D \) denote the set of tuples across all patients in the study. The problem of learning a DSS function is defined by the sets \( O \) and \( S \) of pairs of tuples from the set \( D \) of all tuples, and by the set \( G \) of permissible DSS functions. The set \( O \) contains pairs of tuples \( (< x_i^p, t_i^p >, < x_j^q, t_j^q >) \) that
are ordered by severity based on clinical assessments, i.e., $x_{1}^{p}$ corresponds to a more severe state than $x_{2}^{q}$. We refer to each of these paired tuples as a *clinical comparison* and the set $O$ as the set of all available clinical comparisons. These clinical comparisons can be obtained by presenting clinicians with data $x_{i}^{p}$ for patient $p$ at time $t_{i}^{p}$ and data $x_{i}^{q}$ for patient $q$ at time $t_{i}^{q}$. For each such pair of feature vectors, the clinical expert identifies which of these correspond to a more severe health state; the expert can choose not to provide a comparison for a pair where the severity ordering is ambiguous. These pairs can also be generated in an automated fashion by leveraging existing clinical guidelines. In the Experimental Methods section, we describe how we use an existing guideline in the task of learning sepsis severity.

Below, we describe the L-DSS algorithm for learning linear DSS functions. This algorithm builds on the widely-used soft max-margin training technique that seeks to maximize the distance between the pairs that are at different severity levels. We briefly review the key concepts of soft max-margin ranking before we describe these are adapted to the task of learning a linear DSS function.

**Soft Max-Margin Ranking**: Consider the toy example shown in Figure 1. In this example, we assume that all feature vectors are taken on same patient and the time at which these feature vectors are taken is irrelevant. We simplify the notation accordingly. Let $D$ contain the three feature vectors $\{x_{1}, x_{2}, x_{3}\}$ where $x_{i} \in \mathbb{R}^{2}$, and $O$ contain the pairs $(x_{2}, x_{1})$ and $(x_{3}, x_{2})$, i.e., feature vectors $x_{2}$ and $x_{3}$ have higher disease severity than $x_{1}$ and $x_{2}$ respectively.

Max-margin ranking seeks to find a vector $w$ such that the *margin* between pairs of different severity levels is maximized. In our example, we show parameter vectors $w_{1}$, $w_{2}$ and $w_{3}$ for three candidate ranking functions in Figure 1. For each feature vector $x$, the assigned (severity) score for a given ranking function parameter $w_{i}$ is computed as the projection, $g_{w_{i}}(x)$, of $x$ on $w_{i}$. The induced ranking between two vectors $x_{1}$ and $x_{2}$ is computed based on the margin which is defined as the difference in their projections. In the example shown, the rankings induced by both $g_{w_{1}}$ and $g_{w_{3}}$ correctly order all pairs in $O$, i.e., $g_{w_{1}}(x_{3}) > g_{w_{1}}(x_{2}) > g_{w_{1}}(x_{1})$ and $g_{w_{3}}(x_{3}) > g_{w_{3}}(x_{2}) > g_{w_{3}}(x_{1})$, while the rankings induced by $w_{2}$ do not. Furthermore, $w_{3}$ also induces an ordering with a larger margin between the pairs in $O$. Margin-maximization leads to an ordering that is more robust with respect to noise in $x$. Finally, the set $G$ contains a parameterized family of candidate DSS functions $g$ that map feature vectors $x$ to a scalar severity score. In this paper we focus on linear DSS function alone, however, the DSSL framework extends to non-linear DSS functions as well.

In the DSSL framework, our goal is to learn parameters of the function $g \in G$ that quantifies the severity of the disease state represented by a feature vector $x$. This is done using an empirical risk minimization approach, i.e., an objective function is constructed that maps functions $g \in G$ to their empirical risk. This function objective contains two key terms. The first term penalizes $g$ for pairs of tuples ($< x_{i}^{p}, t_{i}^{p} >$, $< x_{j}^{q}, t_{j}^{q} >$) $\in O$ for which the severity ordering induced by $g$ on vectors $x_{i}^{p}$ and $x_{j}^{q}$ is inconsistent with the ground truth clinical assessment. The second term imposes penalty of high rates of temporal changes of the DSS value, thus encouraging selection of a temporally smooth DSS function. The smoothness requirement allows to learn severity scores that mimic the natural inertia exhibited by biological systems. In what follows we present the formal definition of this objective function for learning a linear DSS.
More formally, for each pair of feature vectors \((x_i, x_j) \in O\), we define the margin of their separation by the function \(g_w(x)\) as \(\mu_{ij}^w = g_w(x_i) - g_w(x_j)\). The maximum-margin approach suggests that we can improve generalization and robustness of the learned separator by selecting \(w\) that maximizes the number of tuples that are ordered correctly (i.e., \(\mu_{ij}^w > 0\)) while simultaneously maximizing the minimal normalized margin \(\frac{\mu_{ij}^w}{\|w\|}\). Using the standard soft margin framework, the SVMRank algorithm\(^{21}\) approximates the above-mentioned problem as the following convex optimization program:

\[
\min_{w, \zeta_{i,j}} \left[ \frac{1}{2} \|w\|^2 + \frac{\lambda_0}{|O|} \sum_{(x_i, x_j) \in O} \zeta_{i,j} \right], \quad \text{subject to the ordering constraints:}
\]

\[
\forall(x_i, x_j) \in O: \quad g_w(x_i) - g_w(x_j) = w^T(x_i - x_j) \geq 1 - \zeta_{i,j} \quad \text{and} \quad \zeta_{i,j} \geq 0
\]

Joachims et al.\(^{21}\) solved this optimization program in the dual formulation\(^{21}\). Chapelle et al.\(^{22,23}\) proposed to transform the problem in (1) into a twice differentiable unconstrained convex problem and solve it in its primal form. To this end, we observe that for every value of \(w\), the optimal values of \(\zeta_{i,j}\) are given by \(\zeta_{i,j}^* = \max\{0, 1 - w^T(x_i - x_j)\}\). Substituting to (1), we obtain the following unconstrained convex optimization problem:

\[
\min_{w} \left[ \frac{1}{2} \|w\|^2 + \frac{\lambda_0}{|O|} \sum_{(x_i, x_j) \in O} \max\{0, 1 - w^T(x_i - x_j)\} \right].
\]

The terms of the form \(\max\{0, a\}\), also called the hinge loss, are not differentiable at \(a = 0\). We approximate these terms with the Huber loss \(L_h\) for \(0 < h < 1\) given by

\[
L_h(a) = \begin{cases} 
0 & \text{if } a < -h \\
(a + h^2)/(4h) & \text{if } |a| \leq h \\
a & \text{if } a > h
\end{cases}
\]

It can be seen that \(L_h(a)\) is identical to hinge loss for \(a < -h\) and \(a > h\). This approximation yields the following unconstrained, convex, twice-differentiable optimization problem:

\[
\min_{w} \left[ \frac{1}{2} \|w\|^2 + \frac{\lambda_0}{|O|} \sum_{(x_i, x_j) \in O} L_h(1 - w^T(x_i - x_j)) \right].
\]

The L-DSS Objective and Optimization Algorithm: We now describe the L-DSS algorithm for learning linear DSS functions. We return to our original setting, where we are given sets \(O\) and \(S\) which contain feature vectors belong to more than one patients at varying times.

The L-DSS objective function is obtained by augmenting the Eq. (2) with the following term:

\[
\sum_{(x_i^p, t_i^p), (x_{i+1}^p, t_{i+1}^p) \in S} \left[ \frac{w^T(x_{i+1}^p - x_i^p)}{t_{i+1}^p - t_i^p} \right]^2.
\]

This term penalizes DSS functions that exhibit large changes in the severity score over short durations, hence encouraging selection of temporally smooth DSS functions. Substitution of this term yields the following L-DSS objection function:

\[
\min_{w} \left[ \frac{1}{2} \|w\|^2 + \frac{\lambda_0}{|O|} \sum_{(x_i^p, t_i^p), (x_{i+1}^p, t_{i+1}^p) \in O} L_h(1 - w^T(x_i^p - x_j^p)) \right],
\]

\[
+ \frac{\lambda_8}{|S|} \sum_{(x_i^p, t_i^p), (x_{i+1}^p, t_{i+1}^p) \in S} \left[ \frac{w^T(x_{i+1}^p - x_i^p)}{t_{i+1}^p - t_i^p} \right]^2
\]

1893
This optimization program is solved using the Newton-Raphson algorithm.

**Experimental Methods**

In this section we derive a sepsis severity score by applying the L-DSS method to the large scale MIMIC-II dataset\(^{19}\). With some abuse of terminology, we refer to the resulting sepsis severity score as the L-DSS-Sepsis score. We assess the quality of the obtained L-DSS-Sepsis score by testing its sensitivity to severity changes of different granularity. In our first experiment, we test whether the L-DSS-Sepsis score is sensitive to significant changes severity. Specifically, we test whether the L-DSS-Sepsis score can accurately distinguish and order correctly by their severity different stages of sepsis. We show that the L-DSS-Sepsis score outperforms on this task two clinical scores that are widely used in the ICU setting. In two additional experiments, we analyze the sensitivity of the L-DSS-Sepsis score to a finer-grained changes in sepsis severity. In particular, in the second experiment, we show that the L-DSS-Sepsis score is sensitive to changes in severity state leading up to septic shock. This suggests that the L-DSS-Sepsis score can potentially be used for early detection of this adverse event. In the third experiment, we show that the L-DSS-Sepsis score is sensitive to post-therapy changes in severity and hence could potentially be used to assess patient’s treatment response.

The rest of this section proceeds as follows. We begin with the brief overview of MIMIC-II, the dataset used for learning the sepsis severity score. We then describe the clinical guideline for coarse grading of sepsis severity and show how this coarse grading can be used for automatic generation of clinical comparison pairs. Next, we describe how to choose the values of free parameters for the L-DSS algorithm. We conclude with the detailed description of the three experiments that are used to assess the quality of the learned L-DSS-Sepsis score.

**Dataset:** We use MIMIC-II, a publicly available dataset containing electronic health record data from patients admitted to the ICUs at the Beth Israel Deaconess Medical Center from 2001 to 2008\(^{19}\). We only include adults (> 15 years old) in our study \((N = 16,234)\). We compute 42 different features that are derived from vital sign measurements, clinical history variables, and laboratory test results.

**Sepsis Severity Grading and Automatic Generation of the Clinical Comparison Pairs:** One of the inputs required by the L-DSS algorithm is the set \(O\) of clinical comparisons. For sepsis, these pairs can be created automatically by leveraging the coarse severity grading of sepsis proposed in the Surviving Sepsis Campaign (SSC)\(^{20}\). SSC classifies sepsis development into the following four stages with decreasing levels of severity: septic shock, severe sepsis, SIRS, and “none” (i.e., none of the above). For each of these stages, the guideline defines criteria as combination of 1) thresholds for individual measurements that capture deviation of this measurement from its normal range, and 2) presence of specific diagnosis codes or diagnoses noted in their clinical notes. However, not all measurements needed for implementing the SSC criteria may be available at the given time. We thus relax the definition of the SSC guidelines so it uses measurements made within a short time window prior to the given time. For instance, for fast changing signals (e.g., blood pressure) we use a two hour window, and for slower changing signals (e.g., creatinine), we use an eight hour window. If no measurements were taken in the designated time window, we consider sepsis stage to be unknown. We implement these criteria and automatically identify tuples in the dataset where these SSC definitions are met.

We leverage the SSC grading of sepsis severity to automatically generate clinical comparisons in the following manner. First, we randomly assign the 16,234 patients in the MIMIC dataset to training (60%) and testing sets (40%). Within the training set, we assign two thirds of the patients to the development set and the remaining third to the validation set. For each of the development, validation and testing sets of patients we generate a separate set \(O\) of clinical comparison pairs. This is done by randomly sampling an equal number of pairs of feature vectors \((x_p^i, x_q^j)\) for each of six combinations of different sepsis stages according to SSC, i.e., shock-severe, shock-SIRS, shock-none, severe-SIRS, severe-none, and SIRS-none. These pairs include feature vectors that are taken from the same patient (i.e., \(p = q\)) or from different patients (i.e., \(p \neq q\)). Overall, we extracted 30,000 clinical comparison pairs for the validation set, and 60,000 clinical comparison pairs for the testing set. For the development set, we extracted several sets \(O\) of different sizes, ranging from 120 pairs to 60,000 pairs. We construct the set \(S\) of smoothness pairs to contain 80,000 randomly sampled pairs \((x_p^i, x_{p+1}^i)\) of consecutive feature vectors.

When generating clinical comparisons by automating clinical rules, a natural question that one might ask is whether the learned scores simply recover these clinical rules and thus yield no generalization beyond the SSC grading. This is not the case for two reasons. First, the guideline criteria rely on information captured in notes and diagnosis codes that are not available to the L-DSS method. Second, the temporal smoothness term constrains the learned scores to be smooth and to generalize beyond grading for the coarse severity stages. This issue is covered in greater detail in
In addition to sets \( \mathcal{O} \) and \( \mathcal{S} \), the required input of the L-DSS algorithm includes the values of free parameters \( \lambda_0 \) and \( \lambda_5 \). To specify these values, we follow the two-step procedure proposed in Dyagilev and Saria\(^{18} \). First, with \( \lambda_5 \) set to 0, we select the value of \( \lambda_0 \) that maximizes the accuracy of ordering of clinical comparison pairs for the validation set of patients. That is, we count the fraction of clinical comparison pairs in the set \( \mathcal{O} \) that are ordered by L-DSS in concordance with the ordering prescribed by the SSC. We refer to this quantity as the severity ordering accuracy or \text{SOA}. For a given \( \lambda_0 \), we choose the value of \( \lambda_5 \) that maximizes the SOA of the resulting score. When there are several such values, we pick to the largest \( \lambda_5 \) among these maximizers. Namely, when there are multiple scores that achieve the maximal SOA, we give preference to the smoothest one.

**Missing data imputation:** Following Hug\(^{24} \), we used a modified Last Observation Carried Forward (LOCF) method to impute missing data in given feature vectors. Specifically, within a given time window \( w \), the last value of the observed signal was used. The window length \( w \) for each signal was set to be the typical sampling frequency of this signal. Specifically, the value of \( w \) was determined by combining clinical knowledge regarding signal’s typical rate of change and manual inspection of the distribution of duration between consecutive measurements of this signal. If no measurements were taken within the specified window, the missing value was replaced with the population mean.

### Experiment 1: Distinguishing between the Severity Stages of Sepsis

In our first experiment, we evaluate whether L-DSS-Sepsis can distinguish and correctly order the different stages of sepsis severity. We consider L-DSS-Sepsis scores learned from a different number of training clinical comparison pairs. For each size of the training set, we report the SOA on the testing set of the obtained L-DSS-Sepsis score. As a baseline for our assessment, we use the severity ordering accuracy of existing clinical scores that are widely used in ICU\(^{1} \). There exists several general purpose severity scores that have been validated to assess illness severity and risk of mortality among septic patients\(^{5,23–27} \). In this paper we focus on two of them. The first one is the Acute Physiology and Chronic Health Evaluation\(^{2} \) (APACHE II) score, which is a widely used scoring system for assessing general (not necessarily sepsis-related) disease severity in hospitalized individuals. The second one is based on the Sequential Organ Failure Assessment (SOFA) score\(^{6} \), which was originally designed to assess per-organ sepsis-related damage severity. Specifically, we use the total SOFA score which is the sum of SOFA scores of all organ systems.

### Experiment 2: Is L-DSS-Sepsis sensitive to changes in severity leading up to adverse events?

In the second experiment, we assess whether L-DSS-Sepsis is sensitive enough to capture changes in severity that can occur over the time period leading up to an adverse event. To this end, we examine the behavior of L-DSS-Sepsis in the 18 hour period prior to septic shock. We consider all patients with septic shock in our data with at least 18 hours of data prior to septic shock onset (\( N = 684 \)). On these patients, we define three time intervals of interest: 1) 6 hours prior to the onset of septic shock; 2) 6–12 hours prior to the onset of the septic shock; 3) 12–18 hours prior to the onset of septic shock. We denote the average values of the learned scores in these intervals by \( \bar{s}_{6-0}, \bar{s}_{6-12}, \) and \( \bar{s}_{12-18} \), respectively. We calculate values of \( \Delta_1 = \bar{s}_{6-0} - \bar{s}_{6-12} \) and \( \Delta_2 = (\bar{s}_{6-0} - \bar{s}_{6-12}) - (\bar{s}_{12-12} - \bar{s}_{12-18}) \) for each patient. Using the standard one-tailed t-test, we assess the p-value \( p_{\text{trend-up}} \) for whether the obtained \( \Delta_1 \) can be observed by chance under the null hypothesis that \( \Delta_1 \) are drawn from a zero mean distribution. Similarly, we assess the p-value \( p_{\text{rate acceleration}} \) for whether the obtained \( \Delta_2 \) can be observed by chance under the null hypothesis that \( \Delta_2 \) are drawn from a zero mean distribution.

### Experiment 3: Is L-DSS-Sepsis sensitive to post-therapy changes in severity?

We now evaluate whether L-DSS-Sepsis is sensitive to changes in severity due to administration of fluid bolus – a treatment used for septic shock\(^{20} \). Towards this, we use the self-controlled case series method. We compare trends exhibited by L-DSS-Sepsis values over the five hour intervals prior to and post administration of fluid bolus. We refer to the trends over these intervals as \( \Delta_{\text{prior}} \) and \( \Delta_{\text{post}} \). The value of \( \Delta_{\text{prior}} \) is computed as the difference between the value of the L-DSS-Sepsis at the time of treatment administration and the mean value of L-DSS-Sepsis over the five hour interval prior to treatment administration. Similarly, the value of \( \Delta_{\text{post}} \) is calculated as the difference between the mean value of L-DSS-Sepsis over the five hour interval after treatment administration and the value of L-DSS-Sepsis at the moment of treatment administration. If the patient is responsive to fluid therapy, then \( \Delta_{\text{treat}} = \Delta_{\text{post}} - \Delta_{\text{prior}} < 0 \), that is, if the L-DSS-Sepsis was trending up prior to treatment administration, we expect this trend to be attenuated or even reversed by the treatment.

We identify cases of fluid administration events related to sepsis using the following criteria: 1) the patient is experiencing SIRS, severe sepsis or septic shock at the time of treatment administration, and 2) the patient is hypotensive (has systolic blood pressure below 100 mm Hg), a commonly used criteria for prescribing fluids in sepsis. To avoid
confounding due to multiple administration of fluids, we restrict our attention to treatment administrations that were not preceded or followed by another fluid bolus administration within a five hour window. This yielded a total of 81 fluid bolus administration events. Employing the one-tailed t-test, we assess the p-value \( p_{\text{treatment response}} \) for whether the observed values of \( \Delta_{\text{treat}} = \Delta_{\text{post}} - \Delta_{\text{prior}} \) can be observed by chance under the null hypothesis that \( \Delta_{\text{treat}} \) are drawn from a zero mean distribution.

Results and Discussion

In our first experiment we assess the ability of the L-DSS-Sepsis score to distinguish and correctly order the different stages of sepsis severity. Specifically, we train the L-DSS-Sepsis score on clinical comparisons sets \( O \) of different sizes and measure the corresponding values of severity ordering accuracy (SOA). We present the results of this experiment in Figure 2(a). As expected, the SOA is an increasing function of the number of training clinical comparisons. We note that the L-DSS algorithm yields highly accurate scores even with relatively small amount of training examples. In particular, the maximal SOA equals 0.855 and is obtained by L-DSS-Sepsis scores trained on 30,000 clinical comparisons or more. The SOA of 0.847, which is 99\% of the maximal SOA, is obtained for 1200 clinical comparisons or more. The routinely used clinical scores — APACHE-II and Total SOFA — yield SOA of 0.654 and 0.601, respectively. They are thus outperformed by all L-DSS-Sepsis scores that are trained on 120 clinical comparisons or more. This performance achieved by the L-DSS-Sepsis is significant from a clinical standpoint as it orders severity states more accurately than the two widely used clinical scores. Moreover, it suggests that the L-DSS algorithm yields high quality severity scores even when trained on a small number of clinical comparisons. Hence, it can be applied to diseases where clinical comparisons need to be manually created by experts, thus their number should be minimal.

![Figure 2](image_url)

**Figure 2:** (a) Severity ordering accuracy (SOA) of the L-DSS-Sepsis scores on the testing set for number of training clinical comparisons. Vertical line marks 1200 training clinical comparisons. (b) Distribution of the L-DSS-Sepsis values for different stages of sepsis. (c) Temporal trajectory of a severity score over the time period leading up to septic shock. The vertical line marks the onset of septic shock. (d) Temporal trajectory of a severity score before and after administration of fluid bolus. The solid vertical line marks the administration of fluid bolus.

Hereafter, we use the L-DSS-Sepsis score trained on 1200 clinical comparisons. In Figure 2(b) we plot the probability density of the L-DSS values at different sepsis severity stages. We observe that, on average, the value of L-DSS score during septic shock is higher than in other less severe stages of sepsis. Since the L-DSS score is temporally smooth, this suggest that the L-DSS trajectory should trend up over the time period leading up to septic shock, i.e, should be sensitive to changes in severity leading up to adverse events. In our second experiment, we verify this hypothesis. Overall, over 67\% (95% confidence interval: 63\% - 70\%) out of 684 observed values of \( \Delta_1 \) and 57\% (95\% confidence interval: 53\%-60\%) out of 684 observed values of \( \Delta_2 \) were positive. The obtained p-values \( p_{\text{trend-up}} < 10^{-21} \) and \( p_{\text{rate acceleration}} < 2.25 \cdot 10^{-2} \) rule out the null hypothesis in favor of the stated hypothesis, that is, the learned scores trend up significantly prior to a septic shock and this trend accelerates over time. As an illustration, in Figure 2(c), we show the L-DSS trajectory on an example patient leading up to the onset of shock. It can be seen that the severity score of this patient exhibits a clear upward trend prior to the septic shock. Early identification of this trend can potentially alert the caregiver about a need for a medical intervention and prevent the impending adverse event.
In our last experiment, we addressed the question of whether the L-DSS score is sensitive to post-therapy changes in severity. Overall, the change of trend $\Delta_{\text{treat}}$ is negative in at least 77% (95% confidence interval: 68% - 86%) out of 81 recorded values of $\Delta_{\text{treat}}$. The obtained p-value $p_{\text{treatment response}} < 5 \cdot 10^{-8}$ rules out the null hypothesis in favor of the stated hypothesis, that is, that L-DSS shows significant response to therapy. As an illustration, in Figure 2(d), we show the L-DSS trajectory on an example patient 10 hours before and after administration of fluid bolus. It can be seen that the severity score of this patient trends up prior to treatment and trends down post-treatment.

**Conclusion**

In this paper we evaluated the feasibility of automatically learning a score (L-DSS-Sepsis) that tracks the severity of sepsis over time. We validated the learned sepsis severity score using electronic health record data obtained from patients admitted to four different ICUs at an academic medical center over a period of eight years. Compared to existing illness severity scores of APACHE-II and SOFA, the L-DSS-Sepsis score was significantly more accurate in distinguishing between the different stages of sepsis. The L-DSS-Sepsis score also showed face validity: its trajectories were temporally smooth and tended to trend upwards in individuals that progressed to septic shock. Furthermore, the L-DSS-Sepsis score also behaved as expected after fluid administration–as the patient’s health improved, the value of the L-DSS-Sepsis score decreased. These experiments suggest that an automated severity tracking score such as L-DSS-Sepsis may enable early interventions and monitor a patient’s responsiveness to therapy.

Our study has several limitations. While the L-DSS-Sepsis score showed desirable behavior in tracking changes in illness severity, additional studies are needed to validate whether it identifies at-risk patients before the caregiver has identified them as such. Similarly, whether the L-DSS-Sepsis score can be used to develop a tool for measuring fluid responsive remains to be studied. Finally, the key advantage of the DSSL framework is that it is less prone to practice pattern variations. Data from more than hospital using on a similar population of patients is needed to further test this hypothesis.

Overall, the discussed results are promising. The DSSL framework is general purpose and can be applied to other disease domains and populations. Furthermore, we are encouraged by the fact that the DSSL framework yields high quality sepsis severity scores even when given as few as 120 clinical comparisons. This suggests that it can be successfully applied to diseases for which coarse severity grading guidelines are not available and thus clinical comparisons need to be obtained from asking experts to annotate.

**References**


Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks

Mehmet Günhan Ertosun, PhD¹ and Daniel L. Rubin, MS, MD¹,²
¹Department of Radiology, Stanford University, Stanford, CA, USA
²Department of Medicine (Biomedical Informatics Research), Stanford University, Stanford, CA, USA

Abstract

Brain glioma is the most common primary malignant brain tumors in adults with different pathologic subtypes: Lower Grade Glioma (LGG) Grade II, Lower Grade Glioma (LGG) Grade III, and Glioblastoma Multiforme (GBM) Grade IV. The survival and treatment options are highly dependent of this glioma grade. We propose a deep learning-based, modular classification pipeline for automated grading of gliomas using digital pathology images. Whole tissue digitized images of pathology slides obtained from The Cancer Genome Atlas (TCGA) were used to train our deep learning modules. Our modular pipeline provides diagnostic quality statistics, such as precision, sensitivity and specificity, of the individual deep learning modules, and (1) facilitates training given the limited and variable data in this domain, (2) enables exploration of different deep learning structures for each module, (3) leads to developing less complex modules that are simpler to analyze, and (4) provides flexibility, permitting use of single modules within the framework or use of other modeling or machine learning applications, such as probabilistic graphical models or support vector machines. Our modular approach helps us meet the requirements of minimum accuracy levels that are demanded by the context of different decision points within a multi-class classification scheme. Convolutional Neural Networks are trained for each module for each sub-task with more than 90% classification accuracies on validation data set, and achieved classification accuracy of 96% for the task of GBM vs LGG classification, 71% for further identifying the grade of LGG into Grade II or Grade III on independent data set coming from new patients from the multi-institutional repository.

Introduction

Gliomas are the most common primary malignant brain tumors in adults ¹. They can occur anywhere in the central nervous system, but primarily occur in the brain and arise in the glial tissue ². While these tumors are typically malignant, some types do not behave consistently in a malignant fashion. Gliomas can be WHO grades I–IV based on malignant behavior ¹.

These tumors are differentiated by pathologists through visual inspection of histopathology slides. There are three histological types of glioma: astrocytoma, oligodendroglioma, and oligoastrocytoma. The nuclei of these histological types have distinct characteristics that pathologists use for morphology-based classification. For instance, nuclei in oligodendrogliomas typically are round in shape, small in size, and have negligible cell-to-cell variability with uniform nuclear texture whereas, whereas nuclei in astrocytomas are elongated and irregular in shape, with an uneven, rough nuclear texture due to the clumping of chromatin ¹¹. Many gliomas contain either mixtures of these nuclei or have intermediate forms. Nuclei with either variable combinations of oligodendroglioma and astrocytoma components or with morphologically ambiguous forms make the accurate and reproducible histopathology classification of gliomas challenging ¹¹.

Beyond histopathology determination, another key task for the pathologist is to determine the grade of the tumor (I to IV). For this task, the pathologist considers morphological features of the tumor in the histopathology slides, including mitosis, nuclear atypia, microvascular proliferation, and necrosis. Grade I tumors have a low proliferative potential, and therefore they are usually cured by surgical resection. Grade II tumors are infiltrative and tend to recur, with patient survival from 5 to 15 years. They have a low level of proliferative activity and usually progress to higher grades of malignancy. Grade III tumors have histological evidence of malignancy, and these tumors exhibit nuclear atypia and brisk mitotic activity. Grade IV gliomas, also known as glioblastoma multiforme (GBM), are the most aggressive cancer subtype, and they are characterized by the presence of microvascular proliferation and pseudopalisading necrosis ⁴.

1899
Histopathology tumor grading is a crucial activity, since Grade I has the highest overall survival \(^2,5\) and Grade IV (GBM) has the poorest overall survival. The Cancer Genome Atlas (TCGA) database includes grades II and III tumors as a set of lower grade glioma (LGG), and higher grade brain tumors in set of glioblastoma multiforme (GBM). Survival probabilities calculated from TCGA data for GBM and LGG are shown in Figure 1, showing that survival is strongly dependent on the tumor grade. Hence, it is very important to differentiate between different grades of glioma in considering the treatment options. A major challenge in determining the grade of these tumors is that there is high inter-reader variability in determining the tumor grade \(^9\). This may be due to the fact that the image features that are used to classify these tumors into grades are not always clear or difficult to reliably determine by different observers. Computerized image analysis can partially overcome these shortcomings, due to its capacity to quantitatively and reproducibly measure histologic structures on a large-scale \(^10\).

A workflow for analysis of quantitative nuclear features in glioblastoma (GBM) was previously described,\(^10,11\) wherein individual nuclei are segmented, and then nuclear features are computed to characterize the segmented nuclei. Each individual nucleus is described using features from four categories which are nuclear morphometry, region texture, intensity, and gradient statistic. Those features are then used to assign a score to each nucleus, and a correlative analysis of the morphological score with treatment response and patient survival is carried out. When the computerized analysis results were compared to a panel of neuropathologists, the computerized analysis provided better discrimination between GBMs with differing degrees of an oligo-component, at least with regard to predicting response to therapy. This suggests that nuclear morphology analysis is a promising approach to facilitating a better understanding and diagnosis of glioma and predicting response to therapy and survival \(^11\).

Deep learning has various closely related definitions or high-level descriptions, one of which is: “A class of machine learning techniques, where many layers of information processing stages in hierarchical supervised architectures are exploited for unsupervised feature learning and for pattern analysis/classification.” \(^13\) The essence of deep learning is to compute hierarchical features or representations of the observational data, where the higher-level features or factors are defined from lower-level ones. The family of deep learning methods have been growing increasingly richer, encompassing those of neural networks, hierarchical probabilistic models, and a variety of unsupervised and supervised feature learning algorithms \(^13\).

Convolutional Neural Networks (CNNs or ConvNets), are a type of discriminative deep architecture in which layers consisting of a convolutional layer and a pooling layer are often stacked up with one on top of another to form a deep model \(^14\). The convolutional layer shares many weights, and the pooling layer subsamples the output of the convolutional layer and reduces the data rate from the layer below. The weight sharing in the convolutional layer, together with appropriately chosen pooling schemes, endows the CNN with some “invariance” properties (e.g., translation invariance). CNNs have been found highly effective and been commonly used in computer vision and image recognition \(^13\). CNNs have the advantage of automatically learning the appropriate features, as opposed to traditional machine learning approaches that use hand-crafted features.

In this work, we study not only GBM (Grade IV), but also LGG (Grade II and Grade III), and we create a classification pipeline to grade histopathological images of glioma. Our work is significant in that this is a basic, yet very strong sub-typing that is strongly associated with patient survival. Our approach does not analyze single nuclei individually with classical hand-crafted machine learning features, but instead it analyzes the nuclei within the image tiles using deep learning and automatically learns the appropriate features. Best to our knowledge, deep learning has not yet been applied to pathology image analysis for glioma.
**Figure 1.** Survival probabilities for GBM (Grade IV), LGG – Grade II, and LGG – Grade II calculated from data from TCGA database

**Materials and Methods**

**Dataset**

We obtained publicly available digital diagnostic whole-slide hematoxylin and eosin (H&E) stained histopathology images (WSI) from The Cancer Genome Atlas (TCGA), which included two types of brain cancer: glioblastoma multiforme (GBM) and lower grade glioma (LGG). The LGG dataset contains Grade II and Grade III tumors.

**Preprocessing**

Each WSI included in the TCGA dataset can exceed 2GB in size\(^{11}\). In order to process these images and resulting data structures, we partitioned these images into tiles. Tiling was also useful to enable parallel processing of the image to further accelerate the preprocessing.

The tile size was 1024x1024 pixels in size at the 20x resolution which is the same size and resolution used for tissue microarrays (TMAs), which can be used by a trained pathologist to make informed opinions about a whole tumor. Tiling the entire slide produces some tiles that contain little or no tissue, and only tiles that contained tissue occupying at least 90% of the tile area were chosen for further analysis\(^{12}\). Tissue is distinguished by hysteresis thresholding on the grayscale and 8-bit depth complemented image, and nuclei are segmented using morphological top-hat filtering and hysteresis thresholding\(^{12}\).

Due to the success of the prior nuclear morphology-based studies of GBM,\(^{10,11}\) we selected to work on images in which we segment the nuclei during the preprocessing stage, but do not crop the individual nuclei, instead leave them within their original positions within the tile to not to lose the nuclei distribution map within that specific tile.

The tiles, on which we segmented the nuclei, are then further tiled to reduce in size to form samples, which we call “e-microbiopsy (electronic micro biopsy) samples,” as shown in Figure 2. This is done to reduce the sample sizes to a scale so that the deep learning networks can be trained on the GPU. Despite the attractive qualities of CNNs, and despite the relative efficiency of their local architecture, they have still been prohibitively expensive to apply in large scale to high-resolution images, and in the end, the network’s size is limited mainly by the amount of memory available on current GPUs and by the amount of training time that we are willing to tolerate\(^{21}\). For this work, we have decided for the size of the input images to the network to be 256x256 by considering the factors of GPU memory capacity, and the training time it would take for the deep learning networks.
Figure 2. Image preprocessing (a) Whole tissue slide, (b) A single tile taken from the original image, (c) one of several electronic microbiopsy samples that are input into the deep learning pipeline, where the nuclei are segmented yet left at their original positions to preserve their inter-nuclei interaction and distributional properties.

Convolutional Neural Networks (CNNs)

CNNs are representatives of the multi-stage Hubel-Wiesel architecture, which extract local features at a high resolution and successively combine these into more complex features at lower resolutions. CNNs consist of two kinds of layers: convolutional layers (C layers), which resemble the simple cells, and pooling layers (P layers), which model the behavior of complex cells. Each convolutional layer performs a discrete 2D convolution operation on its source image with a filter kernel and applies a non-linear transfer function. The pooling layers reduce the size of the input by summarizing neurons from a small spatial neighborhood.

The standard way to model a neuron’s output \( f \) as a function of its input \( x \) is with \( f(x) = \tanh(x) \) or \( f(x) = (1 + e^{-x})^{-1} \). In terms of training time with gradient descent, these saturating nonlinearities are much slower than the non-saturating nonlinearity \( f(x) = \max(0;x) \). The neurons with this nonlinearity are referred as Rectified Linear Units (ReLUs). Deep convolutional neural networks with ReLUs train several times faster than their equivalents with tanh units, hence we are using rectified linear units in this work.

The final layer of the CNN is fully connected to the preceding layer, which is a loss layer. The loss layer drives learning by comparing an output to a target and assigning a cost to be minimized. We use a softmax-loss as the loss layer, so the output of the fully-connected layer acts as input for the softmax classifier. The softmax loss layer computes the multinomial logistic loss of the softmax of its inputs, and it is conceptually identical to a softmax layer followed by a multinomial logistic loss layer, but provides a more numerically stable gradient.

Implementing and Training the CNNs

The CNNs are trained with the back-propagation algorithm. We are using Caffe for implementing and training the CNNs. Caffe trains models by the standard stochastic gradient descent algorithm. The CNNs are discriminatively trained via back-propagation through layers of convolutional filters and other operations such as rectification and pooling. Layers have two key responsibilities for the operation of the network as a whole: a forward pass that takes the inputs and produces the outputs, and a backward pass that takes the gradient with respect to the output, and computes the gradients with respect to the parameters and to the inputs, which are, in turn, back-propagated to earlier layers.

Evaluation

We tested our pipeline on independent set of data from TCGA patients who were held out during development of our method and that were not used during training and validation (our CNNs did not previously see any portion of tissue samples from these patients).

Results

Modular Deep Learning Classification Pipeline

We built a modular deep learning pipeline comprising an ensemble of Convolutional Neural Networks (CNN). We created an ensemble of two CNNs, as shown in Figure 3. The first CNN classifies a histological slide to GBM vs LGG, and the second CNN determines the tumor grade (grade II, grade III) for the LGG cases. By definition, GBM
is a grade IV glial tumor, hence the output of pipeline yields possible grade of Grade II, Grade III or Grade IV.

Figure 3. Modular deep learning pipeline for grading glioma using an ensemble of Convolutional Neural Networks.

Figure 4 (left) shows the structure of the first CNN that is used for GBM vs LGG classification. This CNN has a LeNet-like architecture \(^{14}\) and consists of 8 layers, comprising convolution, pooling, rectified linear unit (ReLU), a fully connected (FC) layer, and finally a softmax layer. The second CNN that is used for LGG grade classification is shown on the Figure 4 (right). This network consists of 19 layers, and it is deeper compared to first CNN, but the deeper layers have a fewer number of kernels. The largest bottleneck to be aware of when selecting a CNN architecture is the available graphics processing unit (GPU) memory.

Figure 4. (Left) The structure of the first CNN that is used for GBM vs LGG classification. (Right) The structure of the second CNN that is used for determination of LGG grade classification.

At the moment, it is still difficult to design a theoretically optimal CNN architecture for a particular image classification task \(^{24}\). Hence, a common practice is evaluating several CNNs with different layer architectures (independently to the later evaluations) in order to find a suitable CNN architecture for that particular classification task. \(^{25}\)

We trained the each individual CNN separately. When it comes to deep learning, since it could easily take several days or even weeks to train a network and there are at least thousands, sometimes millions, of data samples, it is not practical to use methods like leave-one-out cross-validation. A common practice is to have 80%-20% or 70%-30%, sometimes even 50%-50%, split of data \(^{21}\) into training and validation sets, and use the training set to train the network and then validation set for validation and parameter optimization purposes.

For training the first CNN we used a total of 8750 e-microbiopsy samples coming from 22 whole tissue slides coming from 4 different tissue source sites, with a training subset of 6998 and validation subset of 1752 samples (80%-20% split). For training the second CNN module we used a total of 7066 e-microbiopsy samples coming from 22 whole tissue slides coming from 3 different tissue source sites, with a training subset of 5671 and validation subset of 1395 samples. We stop the parameter optimization and training process once we reach a point where training error is less than or equal to 2%, and validation error is less than or equal to 10%. The factors that are taken into account for choosing these numbers include implementing early-stopping, as it is known to combat overfitting and acts as a regularization technique. Figure 5 shows the accuracy plot during the training and validation of this first CNN module for the task of GBM vs LGG classification.
Figure 5. The accuracy curve for training and validation of the first CNN module that is used for GBM vs LGG classification.

Figure 6 shows the outputs of 1st, 3rd and 5th layers of the first CNN module, illustrating how the scale of information extracted changes in different layers of the network. The first layers extract features at a low-level scale, (such as intra-nuclei or single-nuclei level features); as the sample progresses through deeper layers of the network, coarser features are extracted, such as inter-nuclei and distributional properties.

Figure 6. Visualizing an e-microbiopsy sample at different layers of processing in the first CNN: The outputs of 1st, 3rd and 5th layers are shown. Image features that are extracted vary in terms of spatial scale within different layers of the CNN.

Figure 7 shows the training and validation accuracies for the second module for the task of determining the grade of an LGG case.

Figure 7. The accuracy curve for training and validation of the second CNN module that is used for grade classification of an LGG sample.
Figure 8 shows the weights learned by the first layer kernels of second CNN module. This shows that the CNN has learned low-level morphological features, such as edge and arc like structures, as well as colors belonging to the nuclear stains and their opposing colors on the color wheel. Some of these features could recognize the nuclei and parts of nuclei, and some of them could recognize the inter-nuclear space, or the inter-nuclear boundaries.

![Image of CNN weights](image)

**Figure 8.** The weights learned by first twenty four of the first layer kernels of the second CNN module

Figure 9 shows the visualization of the second CNN module when an e-microbiopsy sample is fed into the network. Outputs of the first, fourth, seventh and tenth layers are shown. Low level features are more local, and deeper layers reveal coarser image features through the network’s hierarchy in which higher-level features are learned from lower level features.

![Visualizations of CNN outputs](image)

**Figure 9.** Visualizing an e-microbiopsy sample input to the second CNN: The outputs of 1st, 4th, 7th and 10th layers are shown.

**Testing on Independent Data**

In our evaluation of the first CNN module, in which 100 e-microbiopsy samples were randomly selected from 10 independent test slides that had not been used during training and validation of our models, we obtained a classification accuracy of 96% for the task of GBM vs LGG classification. For the second CNN module, wherein we randomly selected 100 e-microbiopsy samples from 7 independent LGG test slides not used previously, we obtained classification accuracy of 71% for the task of Grade II vs Grade III classification for LGG. Figure 10 shows the confusion matrices and individual diagnostic qualities of these two modules and their respective tasks.

![Confusion matrices](image)

**Figure 10.** Confusion matrices and diagnostic qualities of the modules

**Discussion**

In this work, we developed and applied a deep learning approach to the problem of automated classification of LGG vs. GBM and determination of tumor grade. We did this through the assembly of two modular CNN components, each specialized to the two different classification tasks. Our preliminary results appear to be promising, showing 96% accuracy for distinguishing LGG and GBM and 71% accuracy for distinguishing Grade II and Grade III LGG.
The task of classifying the type and grade of glioma using image features is not unlike other types of automated image feature machine learning problems in which a set of pre-defined features is used to characterize the image and predict the classification label, and our current work could have been approached using pre-defined features, such as nuclear shape, texture, etc. However, a substantial disadvantage of pre-defined features is the need to know those which are most informative in the classification task. Often the best features are not known, and a method of unsupervised feature learning could be advantageous, particularly if abundant data are available.

By not segmenting the individual nuclei into individual images as it was done in previous studies \(^{10,11}\), but keeping the locality and distributional properties of them through the tissue image, our deep learning networks are extracting not only intra-nuclei and single nuclei features, but also inter-nuclei features, such as their density, distribution and interactions. Deep learning methods “let the data speak,” and they discover the relevant features in the data itself, rather than imposing pre-defined features. It could take a long time to train a CNN; however, once CNNs are trained, it is very fast to perform the classification. This is a benefit of using CNNs, as the conventional machine learning approaches with hundreds of pre-defined features could take a much longer time to classify images, compared to a CNN. Also, if the already trained networks are wanted to be updated for some reason, such as accumulation of new data that is to be included for training, the methods of transfer learning and fine tuning would enable the faster updating of the already trained networks without the need of starting a new training session from scratch which would take longer time.

We proposed and explored using an ensemble of Convolutional Neural Networks to form a modular decision pipeline. There are various reasons that we chose to have an ensemble of modular and rather simpler CNNs, as opposed to using a single, but more complex monolithic CNN. First, we would like to be able to gauge measures of the diagnostic quality and statistical measures of the performance of the test, when it comes to decision making in biomedical field. Having such a modular structure enables us to analyze the diagnostic quality of each individual decision step. Second, we opted for having a pipeline with modular units at each decision step, as this would make it easier to analyze each smaller unit independently and help to obtain features that are more intuitive or interpretable, compared to having a single yet very complex unit which could make it difficult to gain intuition regarding its operation. Third, the size of image datasets are generally smaller in the medical domain due to challenges of acquiring many images, and CNNs are difficult to train using small datasets. \(^{15}\) Hence, in this work, we opted for training smaller networks with fewer output classes compared to a single, very deep and complex one, with the outputs representing all possible cases in which we are interested. As the number of output classes increases, the more samples might be needed to train a network well. Fourth, by having modular CNN units, we could explore different network architectures. For example, in our case, in one unit we had a shallower network with higher number of kernels per layer, and another network was deeper with fewer number of kernels per layer. Finally, having such a modular structure could enable users to be able to selectively use specific paths or units along the pipeline. Such modularity would allow us to integrate our single units to probabilistic graphical models, such as Bayesian networks and Markov decision processes, or to combine with other machine learning approaches such as Support Vector Machines (SVM).

The results of our evaluation of accuracy for LGG Grade classification on independent data (71%) is a reasonable preliminary result, but leaves room for improvement. There are several potential reasons for this performance. First, TCGA data vary in terms of coming from multiple centers, each of which processes the tissues with non-uniform protocols for tissue slicing, staining, image acquisition, and timing of the steps. Differences in slide preparation, microscope, and digitizing device between two batches of data may lead to differences in image properties between the two batches, and these differences, called “batch effects,” can bias the performance estimates of predictive models. \(^{16}\) The quality of WSI is also affected by artifacts introduced during image acquisition and batch effects, resulting from variations in experimental protocol. Data quality is especially challenging in collaborative repositories, such as TCGA, where a large amount of high-throughput data is collected from multiple institutions. \(^{16}\) Apart from artifacts introduced during tissue processing and image acquisition, there are variations within the tissue slides, and portions of pathology images could contain non tumor tissue which may or may not be relevant to the diagnosis, so classification accuracy can be further improved by considering a set of e-microbiopsy samples from a slide and making the decision via a majority voting like scheme. Future work in which our methods are applied to more homogenous data (e.g., that from a single instruction) would be interesting to establish the magnitude of these confounding aspects on the accuracy of our results. Such future work could also motivate research efforts to reduce the impact of such variations on methods such as ours.

Another factor that could impair the accuracy of our method in classifying LGG grade is that the distinction between a Grade II vs Grade III glioma is more subtle compared to distinguishing Grade IV from Grades II and Grade III. It
has been noted that significant problems relate to the interpretation of histologic criteria used to classify and grade gliomas.\textsuperscript{17,18} Problems with tumor grading are most significant for intermediate grade tumors, for which patient survival broadly overlaps that associated with lower and higher grade tumors.\textsuperscript{9} Our results also agree with these prior observations.

![Graph showing survival probabilities for GBM (Grade IV), LGG Grade II, and LGG Grade III](image)

**Figure 11.** Survival probabilities for GBM (Grade IV), LGG Grade II, and LGG Grade III, shown alongside the respective CNN modules that perform the classification

In terms of survival, if we look at the plot calculated from the survival data of all samples in TCGA (Figure 11), GBM cases are clearly distinct, with a very low probability of survival. For LGG Grade II and LGG Grade III, even though the main rate curves are separated, the %95 confidence intervals are overlapping. So, from the survival point of view, the accuracies of CNN modules also share similar behavior, similar to previous observations from others that there are challenges in tumor grading in intermediate grade tumors, for which patient survival broadly overlaps.\textsuperscript{9} On the other hand, our modular pipeline has advantages, because each unit can be executed individually (i.e., classifying GBM vs LGG). Though pathologists can likely distinguish LGG and GBM with similar accuracy as our approach, our methods could be useful in quality improvement initiatives, as a “second look,” or teaching applications. Our modular approach lets us meet the requirements of minimum accuracy levels that are demanded by the context of different decision points within a multi-class (i.e. more than 2 classes) classification scheme.

Moreover, to our knowledge, our work is the first to apply deep learning methods to the task of pathology image diagnosis and classification of tumor grade. There are many facets of CNN architectural optimization that we have not yet explored which we plan on pursuing in future work. We will also work on ways to improve our accuracy by exploring extra steps during the preprocessing stage, and also including information from the tissue stroma.

**Conclusion**

We developed a deep learning-based modular pipeline with ensemble of CNNs for the problem of classification and grading of glioma from digital pathology images. Our modular classification pipeline approach has the advantages of having diagnostic quality statistics of individual modules, making it easier to train these models given limited data availability and being able to explore different CNN structures for each module, facilitating relatively smaller hence easier to analyze CNNs, and giving flexibility to be used as single modules or within the framework of other modeling or machine learning applications. Our deep learning based classification modules achieved more than 90% accuracies on validation data set, and our approach produced 96% accuracy for GBM vs LGG classification, and 71% accuracy for LGG Grade I vs LGG Grade II discrimination on an independent data set coming from new patients from a collaborative repository where data is collected from multiple institutions. These results may be improved in future by leveraging our modular approach enabling us to address and optimize the different components of the task (diagnosis of GBM vs LGG, and determination of tumor grade) separately.
Acknowledgements
We acknowledge, and would like to thank to Jocelyn Baker for the discussions and help related to tiling pathology images. We also would like to thank NVIDIA corporation for GPU donation.
Mehmet Günhan Ertoşun is a fellow at Stanford Cancer Imaging Training (SCIT) Program, and acknowledges its support (NIH T32 CA009695).

References
Secondary Use of EHR Timestamp data: Validation and Application for Workflow Optimization

Michelle R. Hribar, PhD², Sarah Read-Brown¹, Leah Reznick, MD¹, Lorinna Lombardi, MD¹, Mansi Parikh, MD¹, Thomas R. Yackel, MD, MPH, MS², Michael F. Chiang, MD, MA¹,²

¹Ophthalmology and ²Medical Informatics & Clinical Epidemiology; OHSU, Portland, OR

Introduction
Electronic health records (EHRs) have potential to improve the quality, efficiency, and cost of health care.¹⁻⁶ The transition from traditional paper-based care to EHRs within both hospitals and ambulatory practices has been aggressively promoted by federal initiatives⁷⁻⁸ and is rapidly transforming the process of health care delivery throughout the United States.⁹⁻¹¹ However, clinicians have raised concerns that EHR implementation has negatively impacted their real-world clinical productivity.¹²⁻¹⁶ For example, at Oregon Health & Science University (OHSU), we have one of the leading biomedical informatics departments in the world and completed a successful EHR implementation in 2006 that received national publicity. Yet we have published studies showing that OHSU ophthalmologists currently see 3-5% fewer patients than before EHR implementation and require >40% additional time for each patient encounter.¹⁷

Approaches toward improving the efficiency of clinical workflow using EHRs would have significant real-world impact. Clinicians are pressured to see more patients in less time for less reimbursement due to persistent concerns about the accessibility and cost of health care.¹⁸,¹⁹ Providers today are facing increased patient loads along with increased encounter times due to EHR use, but do not have guidance or information about how to meet these demands. For example, ophthalmologists typically see 15-30 patients or more in a half-day session, utilize multiple exam rooms simultaneously, work with ancillary staff (e.g., technicians, ophthalmic photographers), and examine patients in multiple stages (e.g., before and after dilation of eyes, before and after ophthalmic imaging studies). This creates enormous challenges in workflow and scheduling, and large variability in operational approaches.²⁰

Patient wait time is a result of pressure on provider time as well as clinic inefficiency; wait time has been shown to affect patient satisfaction as well as create barriers to health care.²¹²². Mathematics, specifically queueing theory, explains waiting by the mismatch of arrival times and service times (time with a physician).²³ This mismatch can be increased by ad-hoc scheduling protocols that artificially increase patient wait time.²⁴,²⁵ Addressing this mismatch using smarter scheduling strategies has potential for improving patient wait time.²⁶ Studying and evaluating appointment scheduling strategies in clinical settings is impractical, however, since patient and provider time is too valuable for experimentation. Empirical models of clinical processes using discrete event simulation (DES) can evaluate potential scheduling strategies effectively before implementing them in clinical settings. DES requires large amounts of workflow timing data—much more than can reasonably be collected using traditional time-motion studies. We believe that data to address these problems is currently available within EHR. One major benefit of EHR systems is that clinical data can be applied for “secondary use” beyond direct provision of clinical care; current efforts have focused on areas such as clinical research, public health, adverse event reporting, and quality assurance.²⁷⁻²⁹ Data mining the EHR data has been used to determine patient no-shows with success, grouping patients in emergency departments (ED)³¹, and for quality assurance in the ED.³²,³³ DES has been used for quality improvement in healthcare, but not for evaluating scheduling strategies based on secondary use EHR data and detailed workflow data.³⁰,³¹,³⁴,³⁵

In this paper, we present the results of using secondary EHR data for modeling clinical workflow in 3 outpatient ophthalmology clinics at OHSU. Ophthalmology is an ideal domain for these studies because it is a high-volume, high-complexity field that combines both medical and surgical practice. Our results show that the secondary use of EHR data for workflow data shows promise; it matches the trends of observed clinic workflows and is available for thousands of patient encounters. Further, workflow data can be used to build simulation models for evaluating scheduling strategies based on patient classification.
Methods

This study was approved by the Institutional Review Board at Oregon Health & Science University (OHSU).

Study Environment

OHSU is a large academic medical center in Portland, Oregon. The ophthalmology department includes over 50 faculty providers, who perform over 90,000 annual outpatient examinations. The department provides primary eye care, and serves as a major referral center in Pacific Northwest and nationally. We selected 3 outpatient clinics to study: 1) pediatric ophthalmology (LR), 2) comprehensive eye care (LL), and 3) glaucoma (MP). These 3 clinics represent the diversity of outpatient care in ophthalmology at OHSU.

Over several years, an institution-wide EHR system (EpicCare; Epic Systems, Madison, WI) was implemented throughout OHSU. This vendor develops software for mid-size and large medical practices, is a market share leader among large hospitals, and has implemented its EHR systems at over 200 hospital systems in the United States. In 2006, all ophthalmologists at OHSU began using this EHR. All ambulatory practice management, clinical documentation, order entry, medication prescribing, and billing tasks are performed using components of the EHR.

Workflow Modeling and Reference Data Collection

Interviews with staff and observations of each of the three clinics were performed to determine the basic patient flow. For the three outpatient ophthalmology clinics we studied, this flow consists of interactions with ancillary staff and the physicians, along with possible dilation time between interactions. Once the workflow was understood, we performed time-motion studies for 3 - 6 half-days at each of the clinics. One to two observers recorded timestamps of physicians and staff as they entered and exited exam rooms; these timestamps were then processed later to determine the duration of time spent in exam rooms with patients. This observational timing data served as our reference data for validation of the EHR timestamp data.

Collection of EHR Timestamp Data

Through preliminary iterative data collection and analysis, we identified a set of EHR timestamp data that represents the different steps of the patient flow for each clinic. The source for this data is the clinical data warehouse for our EHR (EpicCare; Epic Systems, Madison, WI). While these timestamps are specific to OHSU’s implementation in ophthalmology, comparable timestamps are available for other vendors, installations and specialties.

1. **Start and End of Patient Visit: Check-in, check-out and after visit summary printing timestamps.** For some clinics, the check-out process happens after the patient leaves; in this case, the timestamp of the after visit summary printing better represents the end of the exam.

2. **Start and End of Each Provider Interaction: Audit log timestamps.** Timestamps from the audit log can be used to represent the beginning and ending of individual provider interactions during the course of the office visit. In addition to timestamp data, this required data about providers (user IDs), exam rooms (workstation IDs) and patient encounters (encounter IDs) to select the proper context for the timestamps.

3. **Time of Dilation: Structured ophthalmology documentation form.** Eye dilation information is entered in the structured ophthalmology documentation form of our EHR. From this data, we can determine if a patient’s eyes were dilated and at what time.

After obtaining the data through queries, we processed the data and computed the workflow timings from EHR timestamps using R programming language scripts.³⁶

Data Analysis

For each clinic, we compared the timing data collected through observation to the timings computed from the EHR timestamps. We looked at the difference in interaction lengths with staff and physicians, as well as differences in overall summary statistics using the reference timings we observed and the time estimates computed from the EHR timestamps. We also included one year of EHR timing data from each of the 3 ophthalmology physicians to compare trends with the observed clinic days, because the longer-term goal of using EHR data for workflow timings is to generate a large validated dataset.
Simulation Models
We used Arena simulation software to build a model of one clinic’s workflow—the pediatric clinic (LR)—using one year of EHR timestamp data. In this model, we focused on the steps of the exam: the initial exam by ancillary staff, eye dilation, and the ophthalmology physician exam. We evaluated scheduling strategies based on patient classification; we studied the best time to schedule new patients to minimize average patient wait time.

Results
Clinic Workflow Modeling and Reference Data Collection
To collect time-motion workflow data, we observed each of the 3 clinics for 3-6 half-days (162 clinical encounters, 263 interactions between staff and patients or between ophthalmologists and patients). For the clinics we observed, 15-20 patients were typically seen during the half-day clinic session but ophthalmologists at OHSU may see up to 30 patients per half-day. Figure 1 summarizes the workflow of the clinics. Patients have an initial exam with an ancillary staff member—either a certified ophthalmic technician (tech) or an orthoptist. The pediatric clinic (LR) uses both types of ancillary staff; the other clinics use only techs. After the initial exam, patients’ eyes may be dilated before the physician examines them.

Validation of EHR Timestamp Data for Workflow Timing
After extracting EHR timestamps and computing provider interaction lengths, we compared them to those from observations as shown in Tables 1-3. Overall, the time differences at all 3 clinics were similar: 65–69% of the total EHR interaction estimates for each clinic were ≤3 minutes from the observed reference timings, with only 17–24% of the EHR interaction estimates >5 minutes from the observed reference timings. For the comprehensive eye care (LL) and glaucoma (MP) clinics, the EHR timings for physician exams were particularly close to the observed ones, but the tech exams were not. In these clinics, we observed that techs occasionally charted in exam rooms before and after seeing patients.

<table>
<thead>
<tr>
<th>Table 1: Time Difference EHR Timestamp vs. Observed Data (Pediatric-LR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1 Min.</td>
</tr>
<tr>
<td>Orthoptist exam</td>
</tr>
<tr>
<td>Tech Exam</td>
</tr>
<tr>
<td>Physician Exam</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Time Difference EHR Timestamp vs. Observed Data (Comprehensive-LL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1 Min.</td>
</tr>
<tr>
<td>Tech Exam</td>
</tr>
<tr>
<td>Physician Exam</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Time Difference EHR Timestamp vs. Observed Data (Glaucoma-MP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1 Min.</td>
</tr>
<tr>
<td>Tech Exam</td>
</tr>
<tr>
<td>Physician Exam</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Tables 1-3: Time Difference between EHR Timestamps and Observed Data. For each clinic, we calculated the difference between the observed interaction time and the EHR estimates. For all the clinics, 65–69% of all EHR estimates were within 3 minutes of the observed times; at least 76% of the EHR estimates at each clinic were within 5 minutes.
which resulted in EHR timings that inaccurately overestimated the actual interaction times. Conversely, we observed providers interacting with patients before and after using the EHR, which resulted in EHR timings that inaccurately underestimated the actual interaction times.

To validate the use of large amounts of EHR data, we looked at summary statistics for the observed timings, the EHR timings for the observed days and the EHR data for one year for each of the 3 clinics as shown in Figure 2. While there was often significant variance between observed data and EHR timestamp data, the general trends matched closely. The EHR estimates tended to be high for ancillary staff and low for physicians. We presume that this is due to techs performing documentation before and after exams, and due to physicians interacting with patients before and after using the EHR.

We then performed a statistical analysis comparing the EHR timestamps to the observed ones; the results are given in Table 4. Using a Wilcoxon signed rank test, we compared the observed data to their associated EHR timestamps to determine if they were different. The results are given in the “Obs” position for the median difference and the p

<table>
<thead>
<tr>
<th>Difference Between Observed Timings &amp; EHR Timestamps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp-LL</td>
</tr>
<tr>
<td>Tech</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Tech</td>
</tr>
<tr>
<td>Physician</td>
</tr>
<tr>
<td>Glaucoma-MP</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Tech</td>
</tr>
<tr>
<td>Physician</td>
</tr>
<tr>
<td>Pediatrics-LR</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Tech</td>
</tr>
<tr>
<td>Orthoptist</td>
</tr>
<tr>
<td>Physician</td>
</tr>
</tbody>
</table>

Table 4: Results of statistical analysis of differences between EHR timestamps for observed days and 1 yr. EHR timestamps are significantly different for ancillary staff when observed days are compared to 1 year of EHR data.
values in Table 4. Only two sets of timings were statistically significantly different: the tech for Glaucoma (MP) and the physician for Pediatrics (LR). Since we’re interested in using the EHR timings to generate a large dataset, we want to compare the observed timings to one year of EHR timing data. We used the Wilcoxon rank sum test to determine if the 1 year of EHR timing data was significantly different from our observed timings (given in the “1 yr” position for the median difference and p values in Table 4). In this case, the test shows that all ancillary staff’s 1 year of EHR timings are significantly different than the observed timings, but the physicians’ EHR timings were not. This is consistent with the graph shown in Figure 2: the physician’s average 1 year EHR timings are similar to the observed, but the ancillary staff’s are not. We presumed the difference was due to EHR timings including staff documentation time. However, because documentation is part of the workflow, including it in the timing data may still accurately reflect the workflow in simulation models.

Development of Computer Based Models

To investigate the utility of the EHR timing data, we started with creating simulation models for one clinic, the pediatric ophthalmology clinic (LR). Using the 1 year EHR dataset, we built a simulation model of the clinic workflow, ran the simulation 100 times and measured the average total exam time (the amount of time spent with a provider) and the average wait time (Table 5). As shown, the average exam and wait times in the simulated model are similar to those we observed and those we calculated using the EHR timestamps. This suggests simulation models that use EHR timing data provide a reasonable approximation of workflows, even though they didn’t match the observed ancillary staff interactions. We presume that this is because the EHR timings include documentation, which is a part of the clinic workflow.

To better schedule patients, we have performed preliminary studies based on patient classification using the relationship between clinical and demographic factors and visit length using timing data from the EHR. Table 6 shows the mean exam time as determined by EHR timestamps for a simple patient classification: new and returning patients at the pediatric clinic (LR). When we further break these classifications down into child and adult patients, we see that the means and standard deviations change. The pediatric clinic treats adult strabismus patients who weren’t diagnosed or treated as children. The adult patients’ means and standard deviations are considerably larger than those of children. For example, a returning adult patient takes about as long as a new child patient; therefore, classifying patients and representing them accurately in our model is key for meaningful analysis.

We have also studied a scheduling strategy based on patient classification: testing the ideal time to schedule new patients. As we see in Table 6, new patient exams are longer than returning patients and have greater variability.

We compared different ways of scheduling new patients and compared the resulting wait times and clinic length. For each test, we scheduled 15 patients in a 4 hour half day of clinic, 3 of whom were new patients (1 adult, 2 child). This is consistent with the mix of patients normally seen in the pediatric clinic (LR). The “New First” strategy scheduled the new patients first, while the “New Last” strategy scheduled the new patients last. The “New Near Last” strategy scheduled the new patients in the 11th, 12th and 13th slots with two returning patients in the last two slots. The results are summarized in Table 7, suggesting that scheduling the new patients last reduces wait times for the returning patients.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Exam Time (Min.) (Mean ± SD)</th>
<th>Wait Time (Min.) (Mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>18.9 ± 11.1</td>
<td>41.2 ± 29.0</td>
</tr>
<tr>
<td>EHR</td>
<td>24.4 ± 13.0</td>
<td>35.7 ± 26.1</td>
</tr>
<tr>
<td>Simulated</td>
<td>22.9 ± 2.6</td>
<td>42.7 ± 13.4</td>
</tr>
</tbody>
</table>

Table 5: Average exam time and wait time (minutes). Data are displayed for 4 days of observed data, 1 year of EHR time stamping data from a single physician clinic (Pediatrics-LR), and 100 iterations of software simulated data. Exam times for the simulated data are similar to the observed & EHR data, but there are larger differences in the wait times.

<table>
<thead>
<tr>
<th>New patient</th>
<th>Returning patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam Time (Min.) (Mean ± SD)</td>
<td>Exam Time (Min.) (Mean ± SD)</td>
</tr>
<tr>
<td>Child</td>
<td>30.3 ± 15.1</td>
</tr>
<tr>
<td>Adult</td>
<td>36.1 ± 16.9</td>
</tr>
<tr>
<td>All</td>
<td>31.0 ± 15.1</td>
</tr>
</tbody>
</table>

Table 6: Total exam times for different patient classifications for pediatric clinic (LR). From the EHR timestamps, we calculated the total exam times for different patients. As the classification becomes more specific, the means and standard deviation change. For example, the adult patients take more time and have a greater standard deviation than children, which can skew classifications that do not consider age.
time and variability (but also increases the length of the clinic), and that scheduling the new patients near the end of the clinic is the best strategy for minimizing both the average patient wait and clinic length. We note that scheduling the new patients first results in dramatically longer patient wait times. These wait times are longer than when the schedules don’t take into account patient classification, as is the case in Table 5 for our observed data. EHR timing data and initial simulation models. Our later simulation models predict that scheduling new patients last or near to last will drastically improve the wait time over what we observed.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Patient Wait (Mean ± SD)</th>
<th>Clinic Length (Mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New First</td>
<td>45.4 ± 2.7</td>
<td>274 ± 4.7</td>
</tr>
<tr>
<td>New Last</td>
<td>29.3 ± 2.1</td>
<td>293 ± 5.6</td>
</tr>
<tr>
<td>New Near Last</td>
<td>31.0 ± 1.9</td>
<td>282 ± 5.2</td>
</tr>
</tbody>
</table>

Table 7: Average wait times and clinic lengths for different scheduling strategies. This was based on scheduling 15 patients in a half-day clinic session, with 3 new patients (1 adult, 2 child). Scheduling new patients last results in less average wait time than scheduling them first, but it also causes the clinic to be longer. Scheduling them near the end (2 returning patients after) balances wait time and clinic length.

Typically, secondary uses of EHR data have been for clinical research, quality assurance, and public health, rather than for operational purposes. While emergency departments have used EHR timing data for tracking patients and quality assurance, our study focuses on using EHR data for modeling outpatient workflows and studying scheduling strategies. Our study shows that the data needed to study workflow can be mined from the EHR and that it represents general trends of clinic workflow. Patient flow is a concern for all areas of healthcare, in both inpatient and outpatient settings. As patients move through the stages of their care, bottlenecks occur at points where demand for resources (providers, beds, etc.) exceeds availability. Obtaining workflow data using time-motion studies is not feasible for the large amounts of observations necessary for determining variability, and using EHR large-scale timestamp data may be an alternative approach.

2. Timestamp data are less reflective of patient interaction length when EHR use does not coincide with actual clinical workflow. Data from this study show that EHR data contain information about provider interaction time, but only when the systems are being used in a predictable workflow. It is impossible to discern whether the remaining time is spent with a patient or not. We found that the EHR timestamps overestimated interaction lengths with ancillary staff when they charted in the EHR before and after the patient exam. Of the 31 ancillary staff interactions where the difference was greater than 5 minutes (Tables 1–3), 22 of them (71%) were overestimated by the EHR timestamps. Similarly, we found that EHR timestamps underestimated provider exam times when physicians interacted with patients before and after using the EHR during the exam. Of the 10 physician interactions that were more than 5 minutes different, 9 of them were underestimated by the EHR timestamps. Similarly, statistical analysis shows that the ancillary staff’s 1 year EHR timings are different when compared to the observed timing data; this is due to EHR documentation done prior to and following the patient interaction. Since this documentation time is a part of the workflow, the EHR timing data may still be used successfully in simulation models. In addition, we are currently investigating other automated methods for gathering workflow timing data, such as combining EHR data with timing and location tracking data gathered from indoor location services (i.e. iBeacons). These methods have the potential for generating more complete and accurate timing data than EHR timestamps alone.

3. These data may be used for novel activities such as developing simulation models for alternative clinical scheduling and workflow strategies. Extracting workflow timing data from the EHR for large amounts of patient encounters allows us to determine trends for average times and variability, determine probability distributions for the workflow timings and build simulation models based on these distributions. Our initial simulation model accurately predicts the wait time resulting from the variability and patterns of patient arrival and provider exam times. In addition, the EHR data allows us to classify patients; in this case, looking at adult vs. child patients as well as new vs. returning patients for a pediatric ophthalmology clinic. We found that adult patients and new patients had exam times that were longer and had greater variability. The simulation models suggest that scheduling these

Discussion
This study has the following key findings: 1) secondary use of EHR timestamp data is generally accurate for measuring patient interaction length, 2) timestamp data are less reflective of patient interaction length when EHR use does not coincide with actual clinical workflow, and 3) these data may be used for novel activities such as developing simulation models for alternative clinical scheduling and workflow strategies.

1. Secondary use of EHR timestamp data is generally accurate for measuring patient interaction length.

2. Timestamp data are less reflective of patient interaction length when EHR use does not coincide with actual clinical workflow. Data from this study show that EHR data contain information about provider interaction time, but only when the systems are being used in a predictable workflow. It is impossible to discern whether the remaining time is spent with a patient or not. We found that the EHR timestamps overestimated interaction lengths with ancillary staff when they charted in the EHR before and after the patient exam. Of the 31 ancillary staff interactions where the difference was greater than 5 minutes (Tables 1–3), 22 of them (71%) were overestimated by the EHR timestamps. Similarly, we found that EHR timestamps underestimated provider exam times when physicians interacted with patients before and after using the EHR during the exam. Of the 10 physician interactions that were more than 5 minutes different, 9 of them were underestimated by the EHR timestamps. Similarly, statistical analysis shows that the ancillary staff’s 1 year EHR timings are different when compared to the observed timing data; this is due to EHR documentation done prior to and following the patient interaction. Since this documentation time is a part of the workflow, the EHR timing data may still be used successfully in simulation models. In addition, we are currently investigating other automated methods for gathering workflow timing data, such as combining EHR data with timing and location tracking data gathered from indoor location services (i.e. iBeacons). These methods have the potential for generating more complete and accurate timing data than EHR timestamps alone.

3. These data may be used for novel activities such as developing simulation models for alternative clinical scheduling and workflow strategies. Extracting workflow timing data from the EHR for large amounts of patient encounters allows us to determine trends for average times and variability, determine probability distributions for the workflow timings and build simulation models based on these distributions. Our initial simulation model accurately predicts the wait time resulting from the variability and patterns of patient arrival and provider exam times. In addition, the EHR data allows us to classify patients; in this case, looking at adult vs. child patients as well as new vs. returning patients for a pediatric ophthalmology clinic. We found that adult patients and new patients had exam times that were longer and had greater variability. The simulation models suggest that scheduling these
patients later in the clinic helps to minimize the average patient wait time; this is consistent with previous studies that found that high variability patients are better scheduled at the end of the clinic.43

Limitations
There are several limitations to our study. First, in order to find the necessary data in the EHR, we had to discern when the provider was using the EHR during a patient interaction. At OHSU, we have uniquely named workstations in each exam room, which makes identifying patient interactions easier. If laptops were used, it would be much more difficult to determine when a provider was with a patient versus charting in an office. Second, the EHR timestamps do not always capture time spent with patients when the provider is not using the EHR. While we can determine what times the providers are not using the EHR, we cannot pinpoint what they are doing at those times.

Conclusion and Future Directions
The secondary use of EHR data for workflow models and optimization is promising. We have shown that EHR timestamps represent the trends of workflow timing data even if they do not capture all the details of the workflow. Having workflow timing data available for all patient encounters allows for creating simulation models for testing different clinic scheduling strategies. However, because EHR data does not accurately capture the entire workflow, we are looking at ways to automatically collect complete workflow data using tracking devices and indoor location services (e.g. iBeacons45). From this complete data, we hope to build more accurate simulation models for evaluating scheduling strategies for all of the 3 clinics as well as generalize these models for other outpatient clinics. We plan to validate our simulation results by implementing the new scheduling strategies in the clinics and monitoring the effects on patient wait time. Larger studies are needed to validate this approach for general use, but the secondary use of EHR timestamp data has implications for broadening the use of EHRs from a repository of clinical data toward a holistic tool for managing clinical workflow.

References
Longitudinal Analysis of Computerized Alerts for Laboratory Monitoring of Post-liver Transplant Immunosuppressive Care

Jason Jacobs, BA¹, Scott P. Narus, PhD¹², R. Scott Evans, MS, PhD¹², Catherine J. Staes, BSN, MPH, PhD¹

¹ Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah; ² Intermountain Healthcare, Salt Lake City, Utah

Abstract

Post-liver transplant patients require lifelong immunosuppressive care and monitoring. Computerized alerts can aid laboratory monitoring, but it is unknown how the distribution of alerts changes over time. We describe the changes over time of the distribution of computerized alerts for laboratory monitoring of post-liver transplant immunosuppressive care. Data were collected for post-liver transplant patients transplanted and managed at Intermountain Healthcare between 2005 and 2012. Alerts were analyzed based on year triggered, time since transplantation, hospitalization status, alert type, action taken (accepted or rejected), reason given for the action taken, and narrative comments. Alerts for overdue laboratory testing became more prevalent as time since transplantation increased. There is an increased need to support monitoring for overdue laboratory testing as the time since transplantation increases. Alerts should support providers as they monitor the evolving needs of post-transplant patients over time. We identify opportunities for improving laboratory monitoring of post-liver transplant patients.

Introduction

More than 65,000 liver transplant patients are currently living in the United States.¹ These patients require lifelong immunosuppressive care and monitoring to prevent organ rejection, toxicity, and death following this costly and complex procedure. A capitated model may be used to reimburse providers for the lifelong care of post-transplant patients, incentivizing transplant centers to minimize costs.² Achieving higher quality care at lower cost is the daunting challenge facing the United States in the era of the Patient Protection and Affordable Care Act and advancing Meaningful Use legislation.³,⁴ Computerized decision support aids have demonstrated potential to support higher quality health care. Their effectiveness has been shown in areas of relatively simple logic, such as checking for medication interactions or recognizing laboratory tests with out-of-range results.⁵,⁶ When appropriately applied, such clinical decision support (CDS) has been shown to reduce errors, decrease costs, and encourage best practice.⁷,⁸ The Office of the National Coordinator for Health Information Technology (ONC) has emphasized the importance of CDS to optimize health care outcomes and the need for its widespread adoption.⁹ CDS has the potential to improve laboratory monitoring for post-transplant immunosuppressive care.¹⁰

In 2004, a computerized alerting system was implemented at Intermountain Healthcare (IH) to support the laboratory monitoring of post-liver transplant patients.¹¹ While internal laboratory results were already available, a data entry program was created to input information regarding laboratory results from external laboratories as structured data in the EHR.¹² The availability of structured data for all laboratory results allowed alerts to be generated for all post-transplant patients regardless of laboratory used for testing. Alerts were generated for new, out-of-range, or overdue immunosuppression-related laboratory testing using automated rules developed by experts in transplant management. The automated rules employed the same logic found in the protocols used by nurse transplant coordinators for routine laboratory monitoring. Alerts were delivered to an electronic inbox within the EHR and remained until accepted or rejected. Nurse transplant coordinators could view the alert message along with information useful for decision making (e.g. date and value of the laboratory result that triggered the alert, date of and time since liver transplantation, and hospitalization status). Coordinators could accept or reject the alert, select the reason for the action taken, and leave a narrative comment. As of 2015, the liver transplant team at IH has continued to use the data entry system and the CDS alerts developed in 2004 to manage their growing population of over 500 active liver transplant patients. A previous study showed that this system led to significant improvements in the completeness, timeliness, and reduced redundancy of laboratory result reporting.¹¹ A more detailed description of the infrastructure, logic, and alerts delivered are available in previous publications.¹¹,¹³
We found one other study that used CDS to support post-transplant laboratory monitoring. Researchers found evidence that CDS improved clinical outcomes and decreased costs during the first year of post-transplant care. Other transplant centers have expressed an interest in implementing CDS to support the lifelong management of their post-transplant patient population. Yet there are studies indicating that CDS may be disruptive or no longer used by target users after initial implementation. While the initial study at IH analyzed alerts over the first five-month period after implementation, there is a gap in the understanding of how alerts are used by nurse transplant coordinators for laboratory monitoring of post-liver transplant patients over time, particularly as time since transplantation increases.

**Objectives**

In this study, we aimed to describe alerts delivered to nurse transplant coordinators from 2005 to 2012. Our objectives were to a) describe the alerts delivered to nurse transplant coordinators to manage patients after liver transplantation over an eight-year period, b) describe the distribution of the alerts and the time to respond to alerts as time since transplantation increased, and c) identify opportunities for improving alerts in order to improve the management of post-transplant immunosuppressive care.

**Methods**

The liver transplant program at IH performed 776 liver transplant surgeries from January 1, 1988, to December 31, 2012. The study population included patients who received a liver transplant at IH and who were monitored for post-transplant laboratory testing of immunosuppressive care by IH nurse transplant coordinators. We included alerts generated between January 1, 2005, and December 31, 2012, for this study. Since individual patient outcomes were not reported for this study, each transplantation was included for patients who received multiple liver transplants (n=15). We classified patients as lost to follow-up during a time period when there was a gap of 365 days or greater between an alert for overdue tacrolimus laboratory testing and an alert for a new tacrolimus laboratory result. We excluded overdue alerts for patients during the time they were classified as lost to follow-up. Patients became active once an alert for a new laboratory result was received. All alerts were triggered by either new (including out-of-range) or overdue laboratory testing.

Data were extracted from the IH enterprise data warehouse (EDW) with the help of EDW experts and a transplant center data manager. Data included immunosuppression and related laboratory results, triggered alerts, and hospital admission and discharge dates and times.

We analyzed alerts based on the year an alert was triggered, time since transplantation, hospitalization status, alert type, action taken (accepted or rejected), reason given for the action taken, and narrative comments. Correlations among data elements were identified. We described the response time between alert generation and action taken, stratified by time since transplantation. We also described the time between an alert for an overdue laboratory test and an alert for a subsequent new laboratory result (including only the first in a series of alerts for overdue testing), stratified by time since transplantation. Time-based results were summarized using a box-and-whisker plot with the ends of the whiskers representing the lowest value within 1.5 times the interquartile range of the first quartile and the highest value within 1.5 times the interquartile range of the third quartile. Outliers (values outside the whiskers) were not shown.

Institutional Review Boards from Intermountain Healthcare and the University of Utah approved this study.

**Results**

Nurse transplant coordinators received alerts for 564 post-liver transplant patients from January 1, 2005, to December 31, 2012. The number of active patients who received laboratory monitoring grew from 338 in 2005 to 418 in 2012 (Table 1).

From January 1, 2005, to December 31, 2012, there were 124,082 computerized alerts delivered to nurse transplant coordinators for laboratory monitoring of post-liver transplant patients. Coordinators received an average of 42.5 alerts per day over this time period, and all alerts were either accepted or rejected. Nearly all (98.0%) alerts were accepted, and 22.8% of alerts were received while the patient was hospitalized. The most common alerts were triggered by new results for creatinine (41.5%) or tacrolimus (30.4%) or when patients were overdue for tacrolimus (12.1%) or creatinine (9.8%) laboratory testing (Table 2).

While the number of alerts per patient remained fairly stable, the number of alerts per day gradually increased over time (39.0 in 2005, 45.6 in 2012) (Table 1). This paralleled an increase in the number of active patients over the
same period (338 in 2005, 418 in 2012), even though the number of transplantations declined. Alerts for overdue laboratory testing constituted a growing proportion of all alerts over time, increasing from 19.6% in 2005 to 29.3% in 2012. There was not a constant change over time in the proportion of alerts generated while patients were hospitalized. The proportion of alerts that were rejected decreased from 7.0% in 2005 to 0.6% in 2008 and remained below 0.6% per year through 2012 (Table 1).

Table 1. Distribution of patients and immunosuppression management alerts, by year

<table>
<thead>
<tr>
<th>Year</th>
<th>Patients</th>
<th>Transplantations (#)</th>
<th>Total Generated (#)</th>
<th>Alerts per Patient (#)</th>
<th>Alerts per Day (#)</th>
<th>Overdue Alerts (%)</th>
<th>Received while Hospitalized (%)</th>
<th>Rejected (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>38</td>
<td>338</td>
<td>14220</td>
<td>42.1</td>
<td>39.0</td>
<td>19.6</td>
<td>18.1</td>
<td>7.0</td>
</tr>
<tr>
<td>2006</td>
<td>38</td>
<td>357</td>
<td>14344</td>
<td>40.2</td>
<td>39.3</td>
<td>16.5</td>
<td>23.0</td>
<td>4.3</td>
</tr>
<tr>
<td>2007</td>
<td>40</td>
<td>372</td>
<td>14733</td>
<td>39.5</td>
<td>40.4</td>
<td>16.9</td>
<td>28.3</td>
<td>4.5</td>
</tr>
<tr>
<td>2008</td>
<td>40</td>
<td>387</td>
<td>15749</td>
<td>40.7</td>
<td>43.1</td>
<td>19.7</td>
<td>25.9</td>
<td>0.6</td>
</tr>
<tr>
<td>2009</td>
<td>37</td>
<td>403</td>
<td>16518</td>
<td>41.0</td>
<td>45.3</td>
<td>21.3</td>
<td>24.6</td>
<td>0.5</td>
</tr>
<tr>
<td>2010</td>
<td>29</td>
<td>402</td>
<td>16146</td>
<td>40.2</td>
<td>44.2</td>
<td>25.1</td>
<td>19.4</td>
<td>0.1</td>
</tr>
<tr>
<td>2011</td>
<td>29</td>
<td>409</td>
<td>15740</td>
<td>38.5</td>
<td>43.1</td>
<td>25.6</td>
<td>21.6</td>
<td>0.2</td>
</tr>
<tr>
<td>2012</td>
<td>26</td>
<td>418</td>
<td>16632</td>
<td>39.8</td>
<td>45.6</td>
<td>29.3</td>
<td>21.9</td>
<td>0.2</td>
</tr>
<tr>
<td>Total</td>
<td>277</td>
<td></td>
<td>124082</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Among alerts for overdue laboratory testing, few (<1%) alerts were generated while patients were hospitalized but up to half of these were rejected (Table 2). When patients were not hospitalized, 8.3% of overdue creatinine alerts and 8.5% of overdue tacrolimus alerts were rejected. The most common reason for rejecting an alert was due to external laboratory results that were available but had not yet been entered into the EHR (Table 3). For accepted alerts, coordinators most frequently sent a notification letter or indicated that they had previously sought to notify the patient. No single patient constituted more than 1% of alerts for overdue laboratory testing.

Table 2. Description of immunosuppression management alerts generated from January 1, 2005-December 31, 2012

<table>
<thead>
<tr>
<th>Alert Message</th>
<th>Alert Count (%)</th>
<th>% Generated While…</th>
<th>% Rejected While…</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Hospitalized</td>
<td>Not Hospitalized</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hospitalized</td>
</tr>
<tr>
<td>Overdue for tacrolimus testing</td>
<td>15010 (12.1%)</td>
<td>0.2%</td>
<td>99.8%</td>
</tr>
<tr>
<td>Overdue for creatinine testing</td>
<td>12217 (9.8%)</td>
<td>&lt;0.1%</td>
<td>&gt;99.9%</td>
</tr>
<tr>
<td>Creatinine (increased by 0.3 since last result)</td>
<td>3912 (3.2%)</td>
<td>42.5%</td>
<td>57.5%</td>
</tr>
<tr>
<td>Creatinine (increased by 0.3 between three results)</td>
<td>1409 (1.1%)</td>
<td>48.5%</td>
<td>51.5%</td>
</tr>
<tr>
<td>Creatinine (no significant increase)</td>
<td>46139 (37.2%)</td>
<td>30.2%</td>
<td>69.8%</td>
</tr>
<tr>
<td>Tacrolimus (below target range)</td>
<td>25932 (20.9%)</td>
<td>24.6%</td>
<td>75.4%</td>
</tr>
<tr>
<td>Tacrolimus (within target range)</td>
<td>8034 (6.5%)</td>
<td>18.0%</td>
<td>72.0%</td>
</tr>
<tr>
<td>Tacrolimus (above target range)</td>
<td>3791 (3.1%)</td>
<td>27.2%</td>
<td>72.8%</td>
</tr>
<tr>
<td>New cyclosporin A</td>
<td>2806 (2.3%)</td>
<td>28.4%</td>
<td>71.6%</td>
</tr>
<tr>
<td>New sirolimus</td>
<td>232 (0.2%)</td>
<td>13.8%</td>
<td>86.2%</td>
</tr>
<tr>
<td>Potassium (below target range)</td>
<td>2261 (1.8%)</td>
<td>64.5%</td>
<td>35.5%</td>
</tr>
<tr>
<td>Potassium (above target range)</td>
<td>521 (0.4%)</td>
<td>55.7%</td>
<td>44.3%</td>
</tr>
<tr>
<td>Magnesium (below target range within 30 days post-transplant)</td>
<td>1204 (1.0%)</td>
<td>21.3%</td>
<td>78.7%</td>
</tr>
<tr>
<td>Magnesium (below target range)</td>
<td>614 (0.5%)</td>
<td>39.5%</td>
<td>60.5%</td>
</tr>
</tbody>
</table>
Conversely, the proportions of alerts generated for hospitalized patients were greater among alerts for new laboratory results (range: 13.8-64.5%) than among alerts for overdue laboratory testing (Table 2). Whether for hospitalized patients or not, few (range: 0-2.2%) alerts for new laboratory results were rejected. Among actions taken for accepted alerts, 100% required no additional action and 48% indicated that the coordinator reviewed the results if the patient was hospitalized, but for non-hospitalized patients, coordinators responded to accepted alerts in a variety of ways, such as reviewing laboratory results, contacting the patient, consulting the physician, or indicating that no action was required (Table 3).

Table 3. Summary of reasons given for rejecting or accepting immunosuppression management alerts, by hospitalization status when the alert was generated

<table>
<thead>
<tr>
<th>Patient hospitalized when alert generated</th>
<th>Patient not hospitalized when alert generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rejected</td>
<td>Accepted</td>
</tr>
<tr>
<td>Overdue laboratory testing alerts:</td>
<td></td>
</tr>
<tr>
<td>57%: No reason given</td>
<td>83%: Patient previously notified</td>
</tr>
<tr>
<td>43%: Non IHC labs available but not yet entered into EHR</td>
<td>4%: Letter notification</td>
</tr>
<tr>
<td></td>
<td>4%: Patient in hospital</td>
</tr>
<tr>
<td></td>
<td>4%: Phone notification</td>
</tr>
<tr>
<td></td>
<td>4%: Unsuccessful phone call</td>
</tr>
<tr>
<td></td>
<td>73%: Non-IHC Labs available but not yet entered into EHR</td>
</tr>
<tr>
<td></td>
<td>13%: No reason given</td>
</tr>
<tr>
<td></td>
<td>13%: Lab testing interval extended by clinician</td>
</tr>
<tr>
<td></td>
<td>57%: Patient previously notified, waiting for labs</td>
</tr>
<tr>
<td></td>
<td>37%: Letter notification</td>
</tr>
<tr>
<td></td>
<td>3%: Unsuccessful phone call</td>
</tr>
<tr>
<td></td>
<td>2%: Phone notification</td>
</tr>
<tr>
<td></td>
<td>1%: No reason given</td>
</tr>
<tr>
<td></td>
<td>&lt;1%: Left message on messaging system</td>
</tr>
<tr>
<td></td>
<td>&lt;1%: Left message with household contact</td>
</tr>
<tr>
<td></td>
<td>&lt;1%: Spoke with patient</td>
</tr>
<tr>
<td></td>
<td>&lt;1%: In person notification</td>
</tr>
<tr>
<td></td>
<td>&lt;1%: Patient in hospital</td>
</tr>
<tr>
<td>New laboratory testing alerts:</td>
<td></td>
</tr>
<tr>
<td>100%: No reason given</td>
<td>100%: No action required</td>
</tr>
<tr>
<td></td>
<td>48%: Reviewed labs</td>
</tr>
<tr>
<td></td>
<td>3%: No reason given</td>
</tr>
<tr>
<td></td>
<td>&lt;1%: Consulted physician</td>
</tr>
<tr>
<td></td>
<td>&lt;1%: Contacted patient</td>
</tr>
<tr>
<td></td>
<td>&lt;1%: Lab results already seen and acted upon</td>
</tr>
<tr>
<td></td>
<td>76%: No reason given</td>
</tr>
<tr>
<td></td>
<td>29%: Lab data charted incorrectly</td>
</tr>
<tr>
<td></td>
<td>47%: No action required</td>
</tr>
<tr>
<td></td>
<td>42%: Reviewed labs</td>
</tr>
<tr>
<td></td>
<td>39%: Contacted patient</td>
</tr>
<tr>
<td></td>
<td>14%: Consulted physician</td>
</tr>
<tr>
<td></td>
<td>5%: No reason given</td>
</tr>
<tr>
<td></td>
<td>5%: Lab results already seen and acted upon</td>
</tr>
</tbody>
</table>

Among alerts received when patients were not hospitalized, the number of alerts per day and the proportion of overdue alerts increased over time (Table 4), similar to the pattern observed overall in Table 1. However, while the proportion of alerts for low tacrolimus laboratory results remained stable, the proportions for normal and high tacrolimus laboratory results declined over time (normal: 8.9% in 2005, 3.8% in 2012; high: 5.3% in 2005, 1.2% in 2012). Likewise, the proportions of alerts for high creatinine laboratory results remained stable while the proportion for normal creatinine laboratory results decreased (35.3% in 2005, 29.8% in 2012). The rejection rate for overdue laboratory testing decreased from 25% or higher during 2005-2007 (range: 25.0-29.4%) to below 3% during 2008-2012 (range: 0.3-2.5%). The rejection rate for alerts of new laboratory results was 2.0% or lower throughout the study.

As time since transplantation increased, the number of alerts per patient declined from 95 to 21 (excluding 10+ years post-transplant) (Table 5). Likewise, the proportion of alerts received while patients were hospitalized decreased from 53.6% during the first period to 13.5% in the last period. Alerts for overdue laboratory testing constituted a growing proportion of alerts as time since transplantation increased from 2.0% for 0-3 months post-transplant to 44.1% for 10+ years post-transplant. In contrast, the proportion of patients with one or more overdue alerts appeared bimodal, with peaks at 3-4 years post-transplant and at 10+ years post-transplant. The proportion of alerts that were rejected ranged from 0.8% to 4.0% with no pattern as time since transplantation increased.

The response time between alert generation and the action taken for non-hospitalized patients increased with time since transplantation. For alerts of new laboratory results, the median response time increased from 6 to 17 hours...
For alerts of overdue laboratory testing, the median response time increased from 6 to 23 hours (Figure 2) but peaked at 39 hours for patients 1-2 years post-transplant. The median response time for alerts of both new and overdue laboratory testing remained fairly stable from 2-3 years post-transplant and beyond.

**Table 4.** Distribution of immunosuppression management alerts generated while patients were not hospitalized, by year

<table>
<thead>
<tr>
<th>Year</th>
<th>Alerts (#)</th>
<th>Alerts per Day (#)</th>
<th>Overdue Alertsa (%)</th>
<th>New Tacrolimusb (%) L</th>
<th>N</th>
<th>H</th>
<th>New Creatininec (%) *</th>
<th>**</th>
<th>O</th>
<th>Other Alertsd (%)</th>
<th>Rejectede (%) OD</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>11650</td>
<td>31.9</td>
<td>23.9</td>
<td>18.7</td>
<td>8.9</td>
<td>5.3</td>
<td>2.6</td>
<td>0.7</td>
<td>35.3</td>
<td>4.6</td>
<td>29.4</td>
<td>2.0</td>
</tr>
<tr>
<td>2006</td>
<td>11052</td>
<td>30.3</td>
<td>21.4</td>
<td>21.9</td>
<td>8.9</td>
<td>3.9</td>
<td>2.4</td>
<td>0.8</td>
<td>36.1</td>
<td>4.7</td>
<td>25.0</td>
<td>0.3</td>
</tr>
<tr>
<td>2007</td>
<td>10569</td>
<td>29.0</td>
<td>23.5</td>
<td>22.1</td>
<td>8.2</td>
<td>3.4</td>
<td>2.8</td>
<td>1.0</td>
<td>35.0</td>
<td>4.1</td>
<td>26.5</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>2008</td>
<td>11666</td>
<td>32.0</td>
<td>26.6</td>
<td>21.6</td>
<td>6.7</td>
<td>2.9</td>
<td>2.0</td>
<td>0.7</td>
<td>34.3</td>
<td>5.3</td>
<td>2.5</td>
<td>0.2</td>
</tr>
<tr>
<td>2009</td>
<td>12450</td>
<td>34.1</td>
<td>28.2</td>
<td>19.3</td>
<td>7.6</td>
<td>3.1</td>
<td>1.9</td>
<td>0.7</td>
<td>34.6</td>
<td>4.5</td>
<td>2.2</td>
<td>0.1</td>
</tr>
<tr>
<td>2010</td>
<td>13017</td>
<td>35.7</td>
<td>31.1</td>
<td>21.1</td>
<td>6.5</td>
<td>2.1</td>
<td>2.3</td>
<td>0.7</td>
<td>32.8</td>
<td>3.4</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>2011</td>
<td>12345</td>
<td>33.8</td>
<td>32.6</td>
<td>21.3</td>
<td>5.2</td>
<td>1.6</td>
<td>2.3</td>
<td>0.7</td>
<td>32.1</td>
<td>4.1</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>2012</td>
<td>12987</td>
<td>35.6</td>
<td>37.5</td>
<td>18.0</td>
<td>3.8</td>
<td>1.2</td>
<td>2.5</td>
<td>0.7</td>
<td>29.8</td>
<td>6.6</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Total</td>
<td>95736</td>
<td>32.8</td>
<td>28.4</td>
<td>20.4</td>
<td>6.9</td>
<td>2.9</td>
<td>2.4</td>
<td>0.8</td>
<td>33.7</td>
<td>4.7</td>
<td>8.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

a. The combined proportion of alerts indicating overdue laboratory testing for tacrolimus or creatinine.
b. The proportions of alerts that were (L) low, (N) normal, or (H) high compared to the target range for tacrolimus laboratory testing, respectively.
c. The proportions of alerts that were for (*) an increase of 0.3 units between two creatinine results, (**) an increase of 0.3 units between three creatinine results, and (O) all other results for creatinine laboratory testing, respectively.
d. The combined proportion of alerts indicating a new laboratory result for: magnesium, potassium, cyclosporin A, or sirolimus.
e. The proportions of alerts that were rejected for (OD) overdue or (New) new laboratory testing, respectively.

**Table 5.** Distribution of immunosuppression management alerts, by time since transplantation

<table>
<thead>
<tr>
<th>Time since Transplant</th>
<th>Active Patients (#)</th>
<th>Alerts (#)</th>
<th>Alerts per Patient (#)</th>
<th>Overdue Alerts (%)</th>
<th>Patients with ≥1 Overdue Alert (%)</th>
<th>Received while Hospitalized (%)</th>
<th>Rejected (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-3 mo</td>
<td>272</td>
<td>25850</td>
<td>95</td>
<td>2.0</td>
<td>26.5</td>
<td>53.6</td>
<td>0.8</td>
</tr>
<tr>
<td>3-6 mo</td>
<td>271</td>
<td>10130</td>
<td>37</td>
<td>13.4</td>
<td>41.7</td>
<td>30.4</td>
<td>4.0</td>
</tr>
<tr>
<td>6-12 mo</td>
<td>273</td>
<td>10538</td>
<td>39</td>
<td>6.4</td>
<td>46.5</td>
<td>19.9</td>
<td>1.1</td>
</tr>
<tr>
<td>1-2 yr</td>
<td>288</td>
<td>11950</td>
<td>41</td>
<td>13.1</td>
<td>63.9</td>
<td>11.4</td>
<td>3.4</td>
</tr>
<tr>
<td>2-3 yr</td>
<td>280</td>
<td>10283</td>
<td>37</td>
<td>25.1</td>
<td>68.6</td>
<td>15.5</td>
<td>3.4</td>
</tr>
<tr>
<td>3-4 yr</td>
<td>264</td>
<td>8752</td>
<td>33</td>
<td>35.6</td>
<td>75.0</td>
<td>11.9</td>
<td>3.3</td>
</tr>
<tr>
<td>4-5 yr</td>
<td>249</td>
<td>6622</td>
<td>27</td>
<td>24.2</td>
<td>54.2</td>
<td>17.0</td>
<td>1.5</td>
</tr>
<tr>
<td>5-6 yr</td>
<td>231</td>
<td>5092</td>
<td>22</td>
<td>25.0</td>
<td>46.8</td>
<td>12.7</td>
<td>1.2</td>
</tr>
<tr>
<td>6-7 yr</td>
<td>221</td>
<td>4749</td>
<td>21</td>
<td>35.1</td>
<td>50.2</td>
<td>10.7</td>
<td>1.0</td>
</tr>
<tr>
<td>7-8 yr</td>
<td>200</td>
<td>4375</td>
<td>22</td>
<td>33.5</td>
<td>55.5</td>
<td>14.5</td>
<td>1.5</td>
</tr>
<tr>
<td>8-9 yr</td>
<td>193</td>
<td>3941</td>
<td>20</td>
<td>35.9</td>
<td>59.1</td>
<td>13.6</td>
<td>2.6</td>
</tr>
<tr>
<td>9-10 yr</td>
<td>174</td>
<td>3700</td>
<td>21</td>
<td>35.3</td>
<td>60.3</td>
<td>11.4</td>
<td>2.9</td>
</tr>
<tr>
<td>10+ yr</td>
<td>193</td>
<td>18100</td>
<td>94</td>
<td>44.1</td>
<td>82.4</td>
<td>13.5</td>
<td>1.6</td>
</tr>
<tr>
<td>Total</td>
<td>564</td>
<td>124082</td>
<td>21.9</td>
<td>84.9</td>
<td>22.8</td>
<td>22.8</td>
<td>2.0</td>
</tr>
</tbody>
</table>
The time between an alert for overdue tacrolimus laboratory testing and the next alert for a new tacrolimus laboratory result for non-hospitalized patients increased with time since transplantation (Figure 3). The median interval increased from 5.8 to 41.2 days from 0-3 months to 10+ years post-transplant.
Discussion

While a few studies have explored how computerized alerts support laboratory monitoring of patients within the first year after transplantation, no studies have analyzed how these alerts are used by nurse transplant coordinators as time since transplantation increases for patients beyond the first year. Our study shows that the distribution of alerts generated to support the laboratory monitoring of post-liver transplant patients changes over time. As time since transplantation increases, there is a greater need to support the process of monitoring patients who are overdue for laboratory testing. In addition, even though the active patient population continued to grow, there was a decline in the number of new post-liver transplant patients at IH. This shift in the population means that a greater proportion of time must be devoted to monitoring patients who are more prone to overdue laboratory testing. Liver transplantation graft failure rates have continued to improve and patients are surviving longer, further increasing the need to monitor immunosuppressive care, particularly for overdue laboratory testing. Transplant coordinators must juggle the contrasting needs of recently transplanted patients, who require frequent new laboratory testing, and the needs of patients who are several years post-transplant and who receive less frequent laboratory testing but are more prone to being overdue for testing. Thus, computerized alerts should be implemented in a way that supports the evolving needs of managing this patient population.

There was a dramatic decrease in rejected alerts for overdue laboratory testing of non-hospitalized patients from 25% or greater during 2005-2007 to less than 3% during 2008-2012. During the study, 73% of these overdue alerts were rejected due to the availability of laboratory results from external laboratories that had not yet been entered into the EHR. The increasing number of overdue alerts generated over time and the substantial decrease in the proportion that were rejected after 2007 may indicate that additional time dedicated to data entry or implementation of electronic laboratory interfaces were used to improve the integration of laboratory results from external laboratories. A nurse transplant coordinator confirmed that an additional employee had been hired in early 2008, in part to focus on data entry of external laboratory results. The challenge of integrating external laboratory results as structured data into the EHR is a significant barrier to the widespread use of CDS. Considerable effort, both in financial cost and in standards development, continues to be spent to overcome this barrier.

By the end of the study period, the proportion of overdue tacrolimus alerts increased while the proportion of new tacrolimus alerts decreased. In addition, the distribution of specific tacrolimus alerts differed: alerts for low tacrolimus results remained stable, but alerts for normal and high tacrolimus results decreased. Overall, the proportion of alerts for low (20.4%) or high (2.9%) tacrolimus laboratory results readily outnumbered the alerts for
normal (6.9%) tacrolimus results. This is particularly unexpected when there was a decline in the number of new patients and an increase in the proportion of patients who have likely had sufficient time post-transplant for providers to maintain patients within the target range. Patient non-compliance is a possible but unlikely explanation. These unexpected differences may also be explained by a mismatch between the unaltered logic of automated rules that trigger the alerts and revised clinical practice. The protocol for immunosuppression had been revised since the automated rules had been implemented, with a downward shift in the target range. Under the revised protocol, nurse transplant coordinators were receiving alerts for low tacrolimus results that were no longer considered below the target range. The automated rules triggering computerized alerts should be updated when the laboratory monitoring protocol is revised.

When the automated rules were implemented, time since transplantation was the main determinant of the desired target range for immunosuppression. In practice, however, nurse transplant coordinators may also adjust this target range based on certain conditions (e.g. Hepatitis C positive status). The target range for tacrolimus is manually decreased for post-liver transplant patients with these conditions. Coordinators must determine whether a patient is positive for these conditions before knowing whether the alert is valid or should be adjusted. Alerts may be improved by further personalizing the logic based on these conditions.

Alert fatigue among physicians is a well-known unintended consequence of alerting systems. While methods for reducing alert fatigue have been demonstrated, the problem persists. One recommendation to minimize alert fatigue is to provide alerts that are non-interruptive. The alerting system analyzed in this study used non-interruptive alerts, or "notifications," that nurse transplant coordinators viewed in an electronic inbox. This may have contributed to the 100% response rate and the 98% acceptance rate for the alerts received by nurse transplant coordinators. In addition, alerts were delivered in a team-based environment to support transplant patient management and were designed specifically to support this workflow. After ten years of experience with the alerting system, nurse transplant coordinators continue to use the system for patient management.

This observational study has limitations. First, the study included only patients who were transplanted at one institution and who received monitoring of immunosuppressive care from the same institution. This population may not be representative of patients at other transplant centers. Second, our definition of patients who were lost to follow-up may have excluded patients who otherwise would have been included in the study.

Conclusion

As patients progress after liver transplantation, overdue laboratory testing becomes more prevalent. Alerts should be capable of supporting providers as they monitor the evolving needs of post-transplant patients over time. Opportunities exist to further improve computerized alerts by maintaining the logic used by existing alerts and by including additional parameters as transplant clinical management practices advance. Implementation of automated laboratory reporting for a greater proportion of reported laboratory results may further reduce cost and the number of erroneous alerts for overdue laboratory testing.

Acknowledgements

The authors would like to acknowledge National Library of Medicine Training Fellowship grant support (T15LM007124). The authors would also like to acknowledge the contributions of Rosemary DiLauro, RN, for clinical expertise and knowledge of the post-transplant care process flow, and Sean Dow for data acquisition.

References


Examining the Distribution, Modularity, and Community Structure in Article Networks for Systematic Reviews

Xiaonan Ji, MS¹,², Raghu Machiraju, PhD², Alan Ritter, PhD², Po-Yin Yen, RN, PhD¹
¹Department of Biomedical Informatics, ²Department of Computer Science and Engineering, The Ohio State University, Columbus, OH

Abstract
Systematic reviews (SRs) provide high quality evidence for clinical practice, but the article screening process is time and labor intensive. As SRs aim to identify relevant articles with a specific scope, we propose that a pre-defined article relationship, using similarity metrics, could accelerate this process. In this study, we established the article relationship using MEDLINE element similarities and visualized the article network with the Force Atlas layout. We also analyzed the article networks with graph diameter, closeness centrality, and module classes. The results revealed the distribution of articles and found that included articles tended to aggregate together in some module classes, providing further evidence of the existence of strong relationships among included articles. This approach can be utilized to facilitate the articles selection process through early identification of these dominant module classes. We are optimistic that the use of article network visualization can help better SR work prioritization.

Introduction
Systematic reviews (SR) provide a summary of evidence from high quality studies for a specific research question. They are regularly used in health care and for health policy making. Evidence-based Medicine (EBM) relies heavily on the use of synthesized, up-to-date research evidence to make decisions. SRs are considered the highest quality source of evidence for EBM.

SR is commonly conducted by domain experts who are able to draft SR scopes, retrieve relevant citations, assess study quality, and synthesize evidence. Expert researchers first identify the SR scope and research questions, and then generate search strategies to explore related databases (e.g. MEDLINE). The search result is a list of citations, which are usually organized in reference management software (e.g. Endnote, Ref-Works). Before synthesizing relevant evidences, expert researchers need to classify articles based on the title and abstract. Then through the triage (or article selection) process, included articles will proceed to the full-text level. In most SRs, expert researchers include only 2% to 30% of citations at the title and abstract level triage, and 1.6% to 27% gets to be included at the full-text level. In other words, expert researchers spend most of their effort excluding non-relevant or low quality studies. As the classification of articles is one of the most resource and time intensive steps, such workload and resource challenges can limit the tractability of an individual review, the ability to fund a review, and also the ability to respond to new evidence that may require an update to an existing review. To accelerate this process, several machine learning (ML) approaches (i.e. Naïve Bayes and Support Vector Machine) were proposed to facilitate and enhance the title and abstract level triage, abstracts screening, or “article selection”.

As SRs aim to identify, appraise, select and synthesize high quality research evidence relevant to the research questions with a well-designed SR scope, we propose a new approach to pre-define the article relationship with similarity metrics in SR reports. The similarity metric is calculated using several MEDLINE elements, such as title, abstract, MeSH, author, and publication type. We hypothesized that relevant (or included) articles should be more similar to each other than to the excluded ones. We could illustrate such an article relationship as an article network. Each article represents as a node and the relationship (similarity) between two articles represents as an edge connecting them. We hypothesize that with article network visualization, we could detect groups or clusters in the article network, especially for relevant (included) articles in a SR report.

Through our research, we have demonstrated the visualization of article relationships using similarity metrics, and discovered community structures for article networks using 20 completed SR reports. Communities are densely connected groups of vertices, with only sparser connections between groups. Communities reveal a non-trivial internal organization of the network, and allow people to infer special relationships among nodes. Communities have been shown to have significant real-world meaning.

We visualized article networks using a force-directed graph algorithm, which is commonly used for generating network graph where pair-wise geometric distances between the drawn vertices match graph theoretic pairwise
distances. Because of the SR article screening process, articles can be classified into three categories: excluded, half-included (articles included at the title/abstract level, but excluded at the full-text level) and included (articles included at both the title/abstract level and the full-text level). We evaluated the distributions of articles from these three categories (excluded, half-included, and included) using graph diameter, closeness centrality, and module classes (communities). Our research questions included: (1) do included articles tend to aggregate together? (2) with a community detection algorithm, do included articles cluster in the same communities (module classes)?

**Methods**

We used 20 completed SRs: 15 SR reports were produced by the Drug Effectiveness Review Project team (DERP; www.ohsu.edu/drugeffectiveness), the data was made publicly available from Cohen, and 5 SR reports were produced by the Cochrane Collaboration. These 20 SRs were completed by experienced and knowledgeable human expert researchers, with inclusion and exclusion decisions made by at least two expert researchers. Table 1 shows the number and percentage of articles included at 1) abstract level decision and 2) full-text level decision. For instance, the review for ACE Inhibitors has a total of 2544 citations. Based on the abstracts, 183 (7.19%) were included; after full-text reading, 41 (1.61%) were included in the ACE Inhibitor SR report. The final inclusion rates range from 0.55% to 27.04%.

Please note that for SRs from DERP (the first 15 SRs in Table 1), full references in MEDLINE format were downloaded using explicit PubMed Identifiers (PMID). For SRs from Cochrane library (The last 5 SRs in Table 1), references in MEDLINE format were retrieved using the established search strategy on PubMed, references from other databases (i.e. EMBASE/Ovid) were not used.

Again, based on the SR article screening process, articles were classified into three categories: excluded, half-included (articles included at the title/abstract level, but excluded at the full-text level) and included (articles included at both the title/abstract level and the full-text level).

### MEDLINE Elements

In order to establish the relationship among articles, we used MEDLINE elements to create the similarity metrics. MEDLINE elements are the fields in the MEDLINE format, that document the major pieces of information of a publication (article). The MEDLINE display format is used in PubMed MEDLINE records. As the most informative elements, title (TI), abstract (AB) and MeSH (MH) elements are widely used in related work to build feature spaces for ML algorithms. Publication type (PT) is also selected by some studies as it may be a key factor for inclusion or exclusion decisions. In our preliminary work, we found that author information also had some predictive value in the article selection process. Therefore, in this study, we used TI, AB, MH, PT and author (AU) element in our experiments.

### Similarity Score

We calculated the similarity using Cosine similarity. Cosine similarity is widely applied to text mining and measures the cosine of the angle between a pair of vectors. It is a common and efficient measure for text comparison, especially for large datasets. Cosine similarity reflects the degree of similarity based on the presence and frequency of words or terms in each text. For every pair of AUs, PTs and MHs, we simply compared

| Table 1. Twenty SR reports’ total article numbers and rate of inclusion |
|---------------------------------|----------------|----------------|
| **Total** | **Abstract** | **Full-text** |
| **ACE Inhibitors** | 2544 | 183 (7.19%) | 41 (1.61%) |
| **ADHD** | 1051 | 92 (9.68%) | 16 (1.56%) |
| **Antihistamines** | 1120 | 363 (32.41%) | 146 (13.04%) |
| **Beta Blockers** | 2072 | 302 (14.58%) | 42 (2.03%) |
| **Calcium Channel Blockers** | 1218 | 279 (22.91%) | 100 (8.21%) |
| **Estrogens** | 368 | 80 (21.74%) | 80 (21.74%) |
| **NSAIDS** | 393 | 88 (22.39%) | 41 (10.43%) |
| **Opioiods** | 1915 | 48 (2.51%) | 15 (0.78%) |
| **Oral Hypoglycemics** | 503 | 139 (27.63%) | 136 (27.04%) |
| **Proton Pump Inhibitors** | 1333 | 238 (17.85%) | 51 (3.83%) |
| **Skeletal Muscle Relaxants** | 1643 | 34 (2.07%) | 9 (0.55%) |
| **Statins** | 3465 | 173 (4.99%) | 85 (2.45%) |
| **Triptans** | 671 | 218 (32.49%) | 24 (3.58%) |
| **Urinary Incontinence** | 327 | 78 (23.85%) | 40 (12.23%) |
| **Antibiotic** | 412 | 74 (17.96%) | 10 (2.43%) |
| **Antineoplastic** | 1294 | N/A | 19 (1.47%) |
| **Antiretrovirals** | 749 | N/A | 38 (5.07%) |
| **Hearing Loss** | 467 | 13 (2.78%) | 3 (0.64%) |
| **Leukaemia** | 328 | 11 (3.35%) | 7 (2.13%) |
them by exact string matching, because a minor difference may completely alter the outcome. For example, even if two authors’ names are very similar, they may be two different people. However, TI and AB are free text. To calculate the similarity between two TIs and between two ABs, we pre-processed TIs and ABs by removing some common words (such as "the", "is", "are", etc.) that appear frequently in text, stemming each word by the classic Porter Stemmer algorithm. This approach, named alphabetic features, also has been verified to be an effective method to represent an article. The resulting similarity score ranges from 0 to 1 for each element, where 0 indicates independence and 1 means exactly the same. In summary, the similarity score is the equally weighted sum of the MEDLINE element(s) similarity, ranging from 0 to 5.

Network Visualization

Force-directed graph drawing algorithms assign forces among the set of edges and the set of nodes of a graph drawing. Spring-like attractive forces are typically used to attract pairs of endpoints of the graph's edges towards each other, while simultaneously electrical repulsive force are used to separate each pair of nodes. In balanced states, the edges tend to have uniform length, and nodes that are not connected by an edge tend to be drawn further apart. Graphs drawn with these algorithms tend to be aesthetically pleasing, exhibit symmetries, and tend to produce crossing-free layouts for planar graphs.

We used Force Atlas layout to visualize article networks. In this graphic layout, each article was represented as a node and the edge connecting a pair of nodes had an edge weight, which was the similarity score of any paired articles. The initial article network was a complete network with edges. Large complete networks are usually complex and unreadable. To provide a more readable network, we filtered edges with lower similarity scores (in this study, the threshold was 1.0). After the filtering, the largest connected network was generated to represent the relationships among most articles. In the Force Atlas algorithm model, the pair-wise geometric distances between the drawn vertices match the graph theoretic pairwise distances. In our results, similar nodes (articles) tended to aggregate together after implementing the algorithm. Adjacent nodes represented articles with more similarities.

Evaluation Measures

**Graph Diameter and Closeness Centrality**

To mathematically represent the distribution of articles in the force atlas layout and then topologically study the aggregation/clustering of the included articles, we took graph diameter and closeness centrality as the evaluation measures.

In connected graphs, there is a natural distance metric between all pairs of nodes, defined by the length of their shortest paths. Graph diameter is the length of the shortest path between the most distanced nodes. In other words, a graph's diameter is the largest number of vertices needed to travel from one vertex to another when paths which backtrack, detour, or loop are excluded from consideration. A disconnected graph has infinite diameter. As the graph diameter is the maximum eccentricity of any vertex in the graph, it is widely used to measure the topology and concentration/centralization of a graph. A more concentrated graph comes with a smaller diameter. In this study, we examined the graph diameter of the largest connected network as well as the sub-graphs from the subsets of articles (included, half-included, and excluded).

Closeness centrality measures the farness from one node to all other nodes. The more central a node, the lower its total distance to all other nodes. Closeness can be considered as a measure of how long it will take to spread information from a node to all other nodes sequentially. In our study, each node had a closeness centrality value. We used the distribution of closeness centrality as a measure to evaluate the network centralization.

**Modularity**

Modularity (community detection) is a measure of network structure. It was designed to measure the strength of division of a network into modules. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Although a diversity of community detection algorithms have been proposed, the quality of community detection is usually measured by modularity and also some benchmark graphs.

We used modularity to examine resulted communities (also called module classes) in article networks. In this study, the implemented community detection algorithm was a modularity optimization based heuristic method for fast
uncovering of communities in large networks. It was first published in Blondel 2008. This algorithm (also called Louvain method) has several advantages, such as easy implementation, fast computation speed, and the capability to handle large and weighted networks. More importantly, comparing to other methods, Blondel’s algorithm has demonstrated to provide higher quality results for community detection.

While working on several commonly used test-case networks with size ranges from 34 to 118 million, Blondel’s algorithm performs the best in modularity and efficiency while compared to three other algorithms. In addition, this algorithm has also been successfully tested on the commonly applied Girvan and Newman (GN) benchmarks and Lancichinetti and Fortunato and Radicchi (LFR) benchmarks. The performance of Blondel’s algorithm is among the best when considering more than 10 popular community detection algorithms, including Rosvall 2008 and Newman’s works. Blondel’s algorithm has also been applied to some popular social networks with millions of nodes like LinkedIn and Twitter. Therefore, we implemented the Blondel’s algorithm, available in Gephi, to examine the article networks in the SR reports.

Networks Implementation in Gephi

Gephi is an open source software for graph and network analysis. It provides an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. It supports Blondel’s algorithm as the modularity function. We used Gephi version 0.8.2 to examine our article networks. There are three parameters for this modularity function, including resolution, randomization, and weight. We used the original modularity function in Blondel’s algorithm. In Gephi, the parameters were adjusted to use “resolution=1.0” and “randomize”. This is also resulted from the nature of Blondel’s algorithm that a random node order is selected for iteration rather than an unnecessarily fixed order. Although different runs might bring slightly different results, the differences are trivial and the overall structure remains the same. In addition, our article networks used article similarities as edge weights, so the “use weight” option was applied.

Result

We examined 20 SR reports, evaluated their graph diameter, closeness centrality, and modularity, and visualized article networks using Force Atlas layout. To provide a readable article network visualization, edges with a weight less than 1.0 were filtered. After filtering, the largest connected network was generated to represent the relationships among most articles, averaging 92.07% of all articles.

To illustrate the article network graph, we used ADHD report as an example to demonstrate the graphical topology and corresponding evaluation measures. Figure 1 shows the article network graphs for ADHD report. The ADHD report had a total of 851 articles; among which, 84 articles were included at the title/abstract level and 20 articles were included at the full-text level. In Figure 1, nodes were colored by inclusion/exclusion classes. Therefore, 20 nodes (included articles) were labeled in green, 64 nodes (half-included articles) were labeled in yellow, and 767 nodes (excluded articles) were labeled in red. To examine the article distribution, we measured the graph diameter of the largest connected network, which was 10; the graph diameter of the sub-network for included articles (green nodes) was 2; the graph diameter of the sub-network for half-included articles (yellow) was 4; and the graph diameter of the sub-network for excluded articles (red) was 9. The graph diameter demonstrated that included articles were similar to each other.

The distribution of closeness centrality for the largest connected network and sub-networks are shown in Figure 2. The closeness centrality of the largest connected network ranges from 1 to 7 (Figure 2a); but the closeness centrality of the sub-network for included articles, ranges from 1 to 2 (Figure 2b). The higher the closeness centrality score is, the more distance the node is to other nodes (less similar). Only if a node was isolated because of the filtering process (edges weighted <1 were eliminated), the closeness centrality score is 0.

In summary, the sub-network of included articles had a smaller graph diameter and relatively smaller closeness centrality values than the largest connected network. We can conclude that included articles tend to have strong relationships (high similarities in MEDLINE elements). Similarly, the sub-network of the included articles also had a smaller graph diameter and closeness centrality values comparing to the sub-network of the half-included articles.
**Figure 1.** Sample article network graphs from ADHD report. Nodes colored by inclusion/exclusion classes (green= included articles; yellow= half-included articles; red= excluded articles).

**Figure 2.** The distribution of closeness centrality for ADHD report

Vertical axis - Closeness centrality
Horizontal axis - Index of articles

**Figure 2.** The distribution of closeness centrality for ADHD report
In Figure 3, it shows the same network graph as Figure 1, but with a different coloring scheme. Here, nodes are colored based on module classes using Blondel’s algorithm. A total of 73 module classes were found. Eight out of 73 modules contain more than 17 nodes, which is 2% of all articles in the ADHD report. The rest of 65 modules were considered minor communities with only a small amount of nodes or even one single node inside. The top three modules, which contain 31.02%, 18.21% and 17.04% of all articles, were colored in red, green, and yellow (in Figure 3) respectively. With this partition, we observed that 18 out of 20 included articles tended to aggregate in the green module class, with the exception of two articles. In other words, the green module class in Figure 3 covers 90% of included articles. The included articles in this green module class (18/155=11.61%), are much higher than those in the entire network (20/851=2.35%). The findings again confirmed our hypothesis that included articles tend to cluster together because of their high similarities to each other. With this scenario, if the article screening process can begin with this community, we could rapidly identify 90% (18 out of 20) relevant articles.

Figure 3. Sample article network graphs from ADHD report. Nodes colored by module classes. A total of 73 module classes were found.

Graph diameter and closeness centrality

To evaluate the graph centralization of the sub-network for included articles, we calculated its graph diameter and closeness centrality range and compared them with the largest connected network (Table 2). Because Skeletal Muscle Relaxants report has the graph diameter of 1, which is not meaningful for calculation, it was not included in the following analysis due to its fracture network. For the remaining 19 SR reports, the average value of largest connected networks’ diameter was 8.26 (SD = 1.76), while the average graph diameter value of the sub-network for included articles was only 3.84 (SD = 1.54). The results show that the graph diameter of the sub-network of included articles was significantly smaller than that of the largest connected network (paired-t test, p<0.001). And for the closeness centrality range of the largest connected network, the largest one was 1~9 while the smallest one was 1~4. The largest closeness centrality range of the sub-network for included articles was 1~5 while the smallest one was only 1~2.

In summary, the graph diameter of the sub-network for included articles was smaller than the graph diameter of the largest connected network. More than half of the SR reports even have less than a half size of the graph diameter. Besides, the closeness centrality range also shows a smaller value and tighter distribution in the sub-network of included articles comparing to the largest connected network, (1~3 vs. 1~6). The results confirmed that included articles tend to aggregate into a few small areas, and these generate “communities” cover 93.75% included articles in average from 19 SR reports in our study.
Module Classes

Table 3 shows modularity results on module classes after the implementation of modularity function (community detection function) for the 20 SRs. We used the number of modules (M), the number of modules with included articles (Mi), and the number of dominant modules with more than 10% included articles (DMi). We also reported the top two modules, which contain most included articles.

Although a large number of modules are detected for the largest connected network (average M = 86), included articles were only found in a few modules (average Mi = 4), and even less were classified into the dominant modules (average DMi = 2). This implies that included articles have strong similarity which tends to cluster in the same module classes. Moreover, the number of included articles found in Module1 (the top module for each report) provides additional aggregation evidence. In average, Module1 covers 25.5% articles overall, but 67.81% included articles. With Module1 and Module2 together, they total cover 89.03% (67.81% + 21.24% = 89.03%) included articles. Due to the space limitation, we only report the top two modules in Table 3. However, other modules except dominant modules contain very few included articles. In many cases, there was only one included article in a module.

Table 2. Graph diameter and closeness centrality

<table>
<thead>
<tr>
<th></th>
<th>Largest connected network</th>
<th>Sub-network of included articles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GD</td>
<td>CCR</td>
</tr>
<tr>
<td>ACE Inhibitors</td>
<td>7</td>
<td>1~5</td>
</tr>
<tr>
<td>ADHD</td>
<td>10</td>
<td>1~7</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>8</td>
<td>1~6</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>8</td>
<td>1~6</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>8</td>
<td>1~6</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>8</td>
<td>1~6</td>
</tr>
<tr>
<td>Estrogens</td>
<td>7</td>
<td>1~5</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>11</td>
<td>1~9</td>
</tr>
<tr>
<td>Opioids</td>
<td>7</td>
<td>1~5</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>5</td>
<td>1~4</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>8</td>
<td>1~7</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>13</td>
<td>1~9</td>
</tr>
<tr>
<td>Statins</td>
<td>9</td>
<td>1~6</td>
</tr>
<tr>
<td>Triptans</td>
<td>10</td>
<td>1~7</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>8</td>
<td>1~6</td>
</tr>
<tr>
<td>Antibiotic</td>
<td>6</td>
<td>1~5</td>
</tr>
<tr>
<td>Antineoplastic</td>
<td>9</td>
<td>1~7</td>
</tr>
<tr>
<td>Antiretrovirals</td>
<td>6</td>
<td>1~5</td>
</tr>
<tr>
<td>Hearing Loss</td>
<td>11</td>
<td>1~8</td>
</tr>
<tr>
<td>Leukemia</td>
<td>11</td>
<td>1~8</td>
</tr>
<tr>
<td>Average</td>
<td>8.26</td>
<td>1~6</td>
</tr>
</tbody>
</table>

GD = Graph Diameter; CCR = Closeness Centrality Range.
* Significant smaller than the largest connected network (p<0.001)

Table 3. Modularity and detected module classes

<table>
<thead>
<tr>
<th>Modularity</th>
<th>ACE Inhibitors</th>
<th>ADHD</th>
<th>Antihistamines</th>
<th>Atypical Antipsychotics</th>
<th>Beta Blockers</th>
<th>Calcium Channel Blockers</th>
<th>Estrogens</th>
<th>NSAIDS</th>
<th>Opioids</th>
<th>Oral Hypoglycemics</th>
<th>Proton Pump Inhibitors</th>
<th>Skeletal Muscle Relaxants</th>
<th>Statins</th>
<th>Triptans</th>
<th>Urinary Incontinence</th>
<th>Antibiotic</th>
<th>Antineoplastic</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>72</td>
<td>73</td>
<td>43</td>
<td>83</td>
<td>89</td>
<td>68</td>
<td>42</td>
<td>44</td>
<td>90</td>
<td>31</td>
<td>51</td>
<td>322</td>
<td>191</td>
<td>57</td>
<td>25</td>
<td>195</td>
<td></td>
</tr>
<tr>
<td>Mi</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>DMi</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>articles covered</td>
<td>707</td>
<td>159</td>
<td>83</td>
<td>187</td>
<td>400</td>
<td>274</td>
<td>75</td>
<td>78</td>
<td>298</td>
<td>161</td>
<td>449</td>
<td>826</td>
<td>1384</td>
<td>205</td>
<td>17</td>
<td>144</td>
<td></td>
</tr>
<tr>
<td>included articles covered</td>
<td>27.97%</td>
<td>18.68%</td>
<td>26.77%</td>
<td>16.7%</td>
<td>19.26%</td>
<td>22.5%</td>
<td>20.38%</td>
<td>19.85%</td>
<td>15.56%</td>
<td>32.01%</td>
<td>33.68%</td>
<td>50.27%</td>
<td>39.95%</td>
<td>39.95%</td>
<td>10.92%</td>
<td>11.13%</td>
<td></td>
</tr>
<tr>
<td>#</td>
<td>40</td>
<td>18</td>
<td>9</td>
<td>50</td>
<td>26</td>
<td>36</td>
<td>34</td>
<td>32</td>
<td>11</td>
<td>37</td>
<td>37</td>
<td>6</td>
<td>67</td>
<td>17</td>
<td>6</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>%a</td>
<td>97.56%</td>
<td>90%</td>
<td>56.25%</td>
<td>34.25%</td>
<td>61.90%</td>
<td>36%</td>
<td>42.5%</td>
<td>78.05%</td>
<td>73.33%</td>
<td>48.53%</td>
<td>72.55%</td>
<td>66.67%</td>
<td>78.82%</td>
<td>70.83%</td>
<td>5.5%</td>
<td>4.73%</td>
<td></td>
</tr>
<tr>
<td>#</td>
<td>80</td>
<td>10</td>
<td>8</td>
<td>200</td>
<td>476</td>
<td>333</td>
<td>66</td>
<td>42</td>
<td>592</td>
<td>146</td>
<td>405</td>
<td>184</td>
<td>1160</td>
<td>183</td>
<td>42</td>
<td>191</td>
<td></td>
</tr>
<tr>
<td>%b</td>
<td>31.29%</td>
<td>5.88%</td>
<td>23.87%</td>
<td>20.86%</td>
<td>22.97%</td>
<td>27.34%</td>
<td>17.93%</td>
<td>10.69%</td>
<td>30.91%</td>
<td>29.03%</td>
<td>30.34%</td>
<td>11.20%</td>
<td>33.49%</td>
<td>22.5%</td>
<td>28.57%</td>
<td>14.76%</td>
<td></td>
</tr>
</tbody>
</table>

GD = Graph Diameter; CCR = Closeness Centrality Range.
* Significant smaller than the largest connected network (p<0.001)
Table 3. Modularity and detected module classes

<table>
<thead>
<tr>
<th>Modularity</th>
<th>Module1</th>
<th>Module2</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Mi</td>
<td>DMi</td>
</tr>
<tr>
<td>Antiretrovirals*</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Hearing Loss</td>
<td>87</td>
<td>2</td>
</tr>
<tr>
<td>Leukaemia*</td>
<td>92</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td>86</td>
<td>4</td>
</tr>
</tbody>
</table>

M= number of modules; Mi= number of modules that contain included articles; DMi= number of dominant modules with more than 10% included articles; 
* The SR report has only one module that contains included articles. Therefore, there is no Module2.

Discussion

The use of article network visualization for SR

In this study, we used the visualization approach to demonstrate article relationships. While article networks bring a more intuitive view of article distribution, the use of graph measures also provides mathematical support for our hypotheses. The aggregation of included articles confirms the feasibility of utilizing article relationships (similarity) to facilitate the article selection process for SR. We believe the visualization approach can be a powerful tool in assisting SR researchers for article screening. Taking the ADHD report as an example (Figure 3), we can rapidly identify included articles if we start the article screening in the green module. 90% (18 out of 20) included articles could be found earlier before screening in other areas, thus significant workload would likely to be saved.

The article network visualization can be applied to (1) identify multiple key communities when the topic of articles is diversified; (2) assign high priority to communities with relevant articles and screen articles from the closet neighbors; (3) assign low priority to communities that contain several known irrelevant articles to save unnecessary workload; (4) customize the network structure with different similarity calculation for edge weights that align to specific sub-aims of a SR scope, e.g. authorship, publication type, keywords; (5) provide potential knowledge discovery from unexpected module classes. Our future work has planned to develop such an article network visualization application to assist SR.

The limitation of force-directed algorithm

Although the basic force-directed approach performs well for small graphs, the results are poor for graphs with more than one or two thousands vertices. This is mainly resulted from the obstacles to scalability and resolution. To be specific, the minimum vertex separation tends to be very small for large graphs, which leads to unreadable drawings. In addition, the typical force-directed algorithms are considered to have a time complexity equivalent to O(n^3), thus the running time could be very long for large graphs. Due to this limitation, the SRs we applied in our study had no more than 4000 articles. Algorithms and layouts like T-SNE which supports implicit structure and dimensionality reductions could be considered for larger SRs in future works.

Future direction

As most included articles exhibiting a tendency to aggregate to a specific graphic region, we demonstrated the existence of such dominant module classes that cover a high percentage of included articles. Screening articles within these module classes first will likely accelerate the speed of discovering relevant articles. For our future work, we plan to work on early identification of the dominant module classes by integrating SR expert researchers’ knowledge on SR scopes and research questions. With the external information, we are able to generate virtual article(s) representing ideal articles for inclusion. Virtual article(s) are likely to be located in the dominant module classes in the article network.

Conclusion

We demonstrated visualizing article relationships for SR with MEDLINE similarity in force layout. We used measures: graph diameter, closeness centrality, and module classes from the perspective of graph theory, to evaluate the centrality and communities of generated article networks. The sub-networks of included articles have a significantly smaller graph diameter than those of the largest connected network (3.84 vs. 8.26, p<0.001), and a
smaller distribution in closeness centrality (1~3 vs. 1~6). Although a large number of modules were detected for the largest connected network (n=86) after the implementation of modularity function, included articles only cluster into a few communities (n=4), and are found in even fewer dominant communities (n=2). More importantly, the top two modules covered 89.03% included articles.

Since most included articles in our research cohort exhibited a tendency to aggregate to specific regions, early identifying and exploring of these regions will likely accelerate the discovery of relevant articles for a SR. We conclude that while relevant articles in SRs share common features and strong relationships (similarities), the article similarity can be utilized to facilitate the article selection process thus shortening and facilitating the most labor intensive aspect of the SR, which is the most intensive job in SR. Besides, through the visualization of article networks, we demonstrated viewing article relationships in a more intuitive way. We also discovered other advantages of visualizing article networks that are infeasible from the common text comparison approach. For example, if there are multiple major aggregated regions in an article network, a branching of SR scopes may be needed. Integrating article network visualization as a decision support tool in the SR process will enable SR researchers to discover particular patterns or communities; thus accelerating the SR production.

References

A Data Quality Ontology for the Secondary Use of EHR Data

Steven G. Johnson, MS1; Stuart Speedie, PhD, FACMI1; Gyorgy Simon, PhD1; Vipin Kumar, PhD2; Bonnie L. Westra, PhD, RN, FAAN, FACMI1,3

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

Abstract

The secondary use of EHR data for research is expected to improve health outcomes for patients, but the benefits will only be realized if the data in the EHR is of sufficient quality to support these uses. A data quality (DQ) ontology was developed to rigorously define concepts and enable automated computation of data quality measures. The healthcare data quality literature was mined for the important terms used to describe data quality concepts and harmonized into an ontology. Four high-level data quality dimensions (“correctness”, “consistency”, “completeness” and “currency”) categorize 19 lower level measures. The ontology serves as an unambiguous vocabulary, which defines concepts more precisely than natural language; it provides a mechanism to automatically compute data quality measures; and is reusable across domains and use cases. A detailed example is presented to demonstrate its utility. The DQ ontology can make data validation more common and reproducible.

Introduction and Background

The healthcare system in the United States continues to adopt electronic health records (EHR) at a rapid pace.1 The EHR is designed to replace a paper chart and to document and facilitate the delivery of care. Since this electronic data is now much more easily accessed than abstracting from paper charts, it is frequently used for other purposes such as clinical effectiveness research, predictive modeling, population health management and healthcare quality improvement. Secondary use of EHR data is expected to improve health outcomes for patients, but the benefits will only be realized if the data that is captured in the EHR is of sufficient quality to support these secondary uses.2 Investigators have shown that EHR data often contain errors that can impact research results, yet only 24% of clinical studies that use EHR data had a data validation section.3 In order to measure the quality of data there must be an understanding of how the data will be used.4

There is no generally accepted quantitative measure of data quality, but Juran gives an often cited qualitative definition as “…high-quality data are data that are fit for use in their intended operational, decision-making, planning, and strategic roles.”5(p.34-8) Data quality may be adequate when used for one task, but not for another. For example, a higher level of data quality is needed to count the number of diabetic patients with controlled HgA1C than to just count the number of patients. A task refers to concepts in a clinical domain and those concepts are represented by the data. For each task, a set of data quality measures must be developed that determine if the data are adequate to perform the task. The healthcare data quality literature provides terminology and definitions and attempts to organize data quality measures, but there is no general agreement on what these measures should be.6 This terminology-based approach defines measures using natural language, which does not adequately represent the relationships between concepts and is too loosely defined to yield a quantifiable measure of data quality. A better approach is to use an ontology which provides a sufficiently rigorous foundation for concept definitions that enable automated methods for calculating data quality measures.

An ontology is a formal, explicit specification of a shared conceptualization.7 Each concept (also called a “class”) in the ontology has a name, attributes, properties (relations to other concepts) and constraints that must always be true for a concept. The key benefits of defining data quality measures in terms of an ontology are that an ontology is: 1) a specification, written in a formal language and able to represent semantics, 2) a shared vocabulary...
that everyone can use to precisely refer to an aspect of the world, and 3) a sufficiently rigorous specification that can be used for logical inference and computation. An ontology is a logical theory about a part of the world and it defines interrelationships between concepts and axioms that should be true about that world. Automated reasoning can be applied to check internal consistency and make inferences beyond what was explicitly stated in the ontology. This automation eliminates the need for redefining the data quality measures for every task in every domain.

No formal healthcare data quality ontology currently exists, but there is research that examines core data quality concepts. Wang and Strong proposed a framework that consolidates 118 different general data quality characteristics into 20 categories. Kahn proposed a healthcare specific framework using a “fit-for-use” data quality model in which he proposes five high-level dimensions. Liaw performed an extensive literature review looking for commonalities on data quality dimensions. He found consensus on the five most common occurring dimensions were “accuracy”, “completeness”, “consistency”, “correctness” and “timeliness”. While there is some agreement among investigators on these high-level dimensions, there is little agreement or consistency in definitions of more granular data quality concepts such as “validity”, “reliability” and “believability”. In a 2012 paper, Weiskopf defined five high-level dimensions of data quality and listed synonyms for each (Table 1).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td>Accessibility, Accuracy, Availability, Missingness, Omission, Presence, Quality, Rate of recording, Sensitivity, Validity</td>
</tr>
<tr>
<td>Correctness</td>
<td>Accuracy, Corrections made, Errors, Misleading, Positive predictive value, Quality, Validity</td>
</tr>
<tr>
<td>Concordance</td>
<td>Agreement, Consistency, Reliability, Variation</td>
</tr>
<tr>
<td>Plausibility</td>
<td>Accuracy, Believability, Trustworthiness, Validity</td>
</tr>
<tr>
<td>Currency</td>
<td>Recency, Timeliness</td>
</tr>
</tbody>
</table>

Table 1: Weiskopf Five Dimensions of Data Quality with Synonyms

While these dimensions capture orthogonal aspects of data quality, they are defined using natural language descriptions and synonyms. As can be seen from Weiskopf’s descriptions, the same terms may be used multiple times to mean different things (i.e. “Accuracy” occurs 3 times), introducing confusion regarding what aspect of data quality is being described. To provide better conceptual clarity and precision, an ontology is needed.

This paper describes the development of a healthcare data quality ontology (DQ ontology) which provides rigorous definitions and can automate the computation of data quality measures. Given formal ontologies for a clinical domain and for a task, the DQ ontology enables measures to be reused without having to reinvent new data quality assessments for every research project. Ontologies for some clinical domains and tasks already exist and researchers can focus on creating additional ontologies that can be used by the DQ ontology to yield quantified measures. This can make it easier to incorporate data quality validation as a standard component of research results. The DQ ontology was developed from a comprehensive list of data quality terms present in the literature. The terms were organized into an ontology and constraints were defined that precisely describe a data quality measure better than natural language and enable quantification of the measure. It makes explicit which data quality concepts depend on the use of the data and which depend on the clinical domain. A detailed example demonstrates the utility of this ontology for quantifying measures and for discussing aspects of data quality.

Materials and Methods

There are a number of methodologies for developing an ontology, but the method described by Noy and McGuiness was selected due to its simplicity and effectiveness. This methodology advocates a seven-step process that takes a list of terms and definitions and turns them into a formal ontology. The first step is to define the scope of the ontology. For this study, the scope is a shared vocabulary of data quality concepts with formal definitions that are automatically computable to quantify data quality. The software development community has had success adopting the approach of a common vocabulary to allow researchers to spend less time defining concepts and more time applying it in research. Next, the reuse of existing ontologies was considered. No formal healthcare data quality ontology exists; but ontologies that describe clinical domains and tasks do exist and will be reused and referenced by the DQ ontology.

In order to enumerate the important terms in the ontology, an extensive PubMed search for articles published between January 1995 and January 2015 was performed to obtain a comprehensive list of terms and definitions that are used to describe healthcare data quality. The goal was to find literature reviews and meta-analyses of papers about healthcare data quality to identify as many core concepts as possible. Also, all articles about informal...
healthcare data quality frameworks or ontologies were examined for key terms and definitions. Keywords included in the query were: (“data quality”) and (“health” or EHR) and (“literature review” or framework or ontology or assessment or model) and (dimensions or accuracy or consistency or completeness or correctness).

There were 181 articles identified, which were manually reviewed by the first author and narrowed to five meta-analyses from Liaw6, Weiskopf34, Kahn11, Chen18, and Lima19. These papers were either reviews of other papers about healthcare data quality or they proposed an informal data quality framework. They all attempted to categorize data quality concepts into semi-orthogonal dimensions. The references from these papers were also reviewed, which yielded an additional five sources: Wang10, Wand29, Chan11, CIHI22, Stvilia23. Collectively, these 10 meta-analyses reviewed 412 papers looking for common aspects of healthcare data quality. There was similarity on high-level concepts such as “correctness”, “consistency” and “completeness”, but there were limited definitions for important terms such as “dataset”, “data”, “measurement”, “metric” and “measure”. Additional papers from the information science literature were found to further define these important concepts24–26.

Ontologies can be specified using a number of methods including OWL27, first order logic, or as UML28. For this paper, the ontology is documented using a UML diagram and a table that lists constraints. A bottom-up approach was taken in which terms and definitions from the meta-analyses were matched and harmonized into equivalent concepts and these concepts where grouped into higher-level categories. Each concept has properties and relationships with other concepts that were discerned from reading the description in the articles. The cardinality of relationships was also defined. Cardinality indicates whether an associated concept is optional, must always occur, or can occur multiple times. For example, a patient must always have a gender, but a blood pressure reading is an optional observation. Constraints were also defined for each concept, describing what should always be true for a concept. The constraints evaluate to a Boolean (true/false) result and can be written in a number of languages including, Object Constraint Language (OCL), first order predicate logic (FOPL), pseudo-code or openEHR constraint language.8,29 For this study, pseudo-code was chosen because it succinctly captures the important aspects of the constraint without introducing a specific, complex syntax.

Results

There were 96 terms and definitions extracted from the literature as a basis for the data quality measures of the ontology. Terms that described the same concept were matched based on their definition and use within the articles. Concepts that appeared in less than three of the articles were deemed non-core and were left out of this version of the DQ ontology. The resulting data quality ontology is shown in Figure 1 as a UML diagram depicting the relationships, attributes, and cardinality of the concepts. For readability, the 19 lower-level Measures were not included in the diagram and are listed in Table 2, which also provides a definition of the measure and a reference to equivalent terms from the meta-analyses. A bold font is used to indicate that a term refers to a concept from an ontology.

The meta-analyses articles make pervasive reference to concepts such as “data”, “information” and “value”. In the DQ ontology, a more precise concept, Representation, defines the lowest level, atomic piece of information that exists in the data being assessed (synonyms for this concept are data field, observation, value, etc). Representations have a DataValue (the part that is stored somewhere) as well as a ValueType that specifies a format to which the DataValue must conform (i.e. numeric quantity, string, choice field, etc). ValueTypes put constraints on the DataValue of the Representation, and can only refer to intrinsic information about the value itself and not to relationships with other Representations. Formal semantics about concepts represented in the data are defined in a separate Domain ontology. Representations have an attribute, DomainConcept, which maps data to a concept in the clinical Domain ontology. There can be multiple Representations for each concept in the Domain. For example, a systolic blood pressure value can be represented as a single number (i.e. 123) or it can be encoded as the first part of a string (i.e. “123/92”). DomainConcepts can also have multiple synonyms in the Domain ontology (i.e. “BP” and “Blood Pressure”), but for the purpose of assessing data quality, they can all be mapped to a single primary DomainConcept (i.e. ‘Blood Pressure’). The Task designates the context or the specific use of the data and is necessary for assessing fitness-for-purpose. The Domain and Task are separate, formal ontologies to which the DQ ontology refers. A Dataset is an arbitrary grouping of Representations of interest. For example, a Dataset can be all of the Representations in the entire EHR.

One of the key concepts in the DQ ontology is the Measure, which is defined as “a quantity that characterizes a quality of the data”. Other possible terms considered were “dimension”, “aspect”, “measurement”, “metric”. Measure was chosen because it captured the notion of quantifying an aspect of interest. The word is used as a noun, not a verb. A Measure is quantified using a MeasurementMethod. A Measurement is a process that performs a MeasurementMethod on a specific Representation (or Dataset) at a point in time that yields a MeasurementResult which is a quantity, usually numeric (but possibly a boolean or text value). A Metric is a
statistic about a series of MeasurementResults along a dimension such as time or across patients. For example, a MeasurementResult could indicate that there were 72 data format errors in a Dataset. But a Metric for that situation would be that there were an average of 5.5 data format errors per day or per patient. This part of the DQ ontology was based in part on core concepts from the Ontology for Software Measurement\textsuperscript{24}.

Four high-level data quality dimensions (CorrectnessMeasure, ConsistencyMeasure, CompletenessMeasure and CurrencyMeasure) categorize 19 lower level Measures. “Accuracy” is one of the terms that had many definitions in the literature. In Weiskopf\textsuperscript{13}, she lists at least 3 different ways that the term is used. It sometimes means only correctness but it is also used to represent completeness or plausibility. For that reason, the term “accuracy” has been avoided in the DQ ontology because it is too overloaded. Instead, the term “correctness” was selected to represent this core concept.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{data_quality_ontology.png}
\caption{Data Quality Ontology}
\end{figure}

1940
<table>
<thead>
<tr>
<th>Concept</th>
<th>Definition</th>
<th>References / Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>CorrectnessMeasure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RepresentationIntegrity</td>
<td>Aspects of the Representation that reassure that data was not corrupted or subject to data entry errors.</td>
<td>Correctness: Credibility of source(^6), Accuracy: ...free of error(^1), Integrity(^8), Repeatability(^9), Structural Consistency(^23)</td>
</tr>
<tr>
<td>RelativeCorrectness</td>
<td>Assesses the quality of a Representation by comparing it to its counterpart in another Dataset which is a &quot;relative standard&quot;, computed as PPV.</td>
<td>Accuracy: ...conformity with actual value(^6), Correctness(^3), Believability(^11), Validity(^12,13,19), Comparability(^10,21), Accuracy(^10,13,18,23,19), Corrections made(^13), Errors(^13), Misleading(^11), PPV(^19), Quality(^13)</td>
</tr>
<tr>
<td>RepresentationCorrectness</td>
<td>A correct Representation has high accuracy and is complete.</td>
<td>Correctness: ...accuracy and completeness(^6), Accuracy(^20,21)</td>
</tr>
<tr>
<td>Reliability</td>
<td>The data is correct and suitable for the Task.</td>
<td>Reliability(^6,18,20), Accuracy: Measurement Error(^22)</td>
</tr>
<tr>
<td>ConsistencyMeasure</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| RepresentationConsistency      | The data is a valid value and format for its DataValueType and all of the Representations for the same information have the same values. | Consistency: ...values and physical representation of data\(^9\), Concordance\(^13\), Format\(^1\), Internal Consistency\(^7\), Consistency\(^9\), Precision\(^7\), Format\(^12\), Reliability\(^3\), Variation\(^1\), Accuracy: Edit and Imputation\(^7\), Representational Consistency\
| DomainConsistency             | Concepts in the Domain are represented in the data and the data satisfies syntactic and semantic rules. Constraints for the Domain are satisfied. | Accuracy: Refers to values and representation\(^6\), Correctness: ...format and types are valid\(^9\), Plausibility\(^9\), Believability\(^11\), Relational Integrity Rules\(^11\), Consistency\(^18\,20\), Measure validity\(^21\), Accuracy\(^13\), Trustworthiness\(^13\), Validity\(^3\,22\), Interpretability\(^10\) |
| CodingConsistency             | Representations that are of coded text data type must be correctly mapped to an enumerated list or a terminology. | Consistency: ...codes/terms...mapped to a reference terminology\(^6\), Valid values\(^11\), Comparability: Equivalency\(^2\), Semantic Consistency\(^22\) |
| DomainMetadata                | Meta-data exists to describe the Domain and it is logically consistent.   | Methodological Clarity\(^13\), Metadata Documentation\(^18\), Comparability: Data dictionary standards\(^2\), Interpretability\(^10\) |
| CompletenessMeasure           |                                                                           |                                        |
| RepresentationComplete        | Domain independent extent to which data is not missing.                   | Completeness: ...information is not missing\(^6\), Completion\(^19\), Completeness\(^1\), Accuracy: Item Non-Response\(^2\) |
| DomainComplete                | The extent to which information is present or absent as expected.         | Appropriate amount of data: Data are present or absent as expected\(^6\), Optionality\(^1\), Content\(^20\) |
| RelativeCompleteness          | The extent to which a truth about the world is represented in the data. This is computed as sensitivity relative to another Dataset. | Completeness: Is a truth...in the EHR?\(^13\), Accessibility\(^10\,13\,15\), Accuracy\(^3\), Availability\(^13\), Missingness\(^13\), Omission\(^13\), Presence\(^13\), Quality\(^13\), Rate of Recording\(^13\), Sensitivity\(^13\), Validity\(^13\) |
| Sufficiency                   | The data has sufficient Representations along a given dimension (i.e. time, patient, encounter) to perform the Task. | Completeness: ...sufficient breadth and depth for the task\(^11\), Appropriate amount of data\(^11\), Representativeness\(^10\), Sufficiency\(^20\), Accuracy: Coverage\(^22\), Granularity\(^11\,18\), Continuity\(^1\), Level of Detail\(^20\), Completeness\(^1\), Precision\(^13\) |
| DomainCoverage                | The data can represent the values and concepts required by the Domain.     | Completeness: ...represent every meaningful state of the [...] real world\(^9\), Completeness: All values for a variable are recorded\(^3\), Coverage\(^19\), Completeness\(^20\) |
| TaskCoverage                  | The data contains all of the information required by the Task.             | Completeness: ...depict every possible state of the task\(^11\), Usability\(^13\), Usability\(^1\), Utility\(^16\), Importance\(^23\), Usefulness\(^23\), Value-added\(^23\) |
| Flexibility                   | The extent to which the data is sufficient to be used by many Tasks.      | Consistency: ...information...applies to different tasks\(^6\), Flexibility\(^1\), Relevance: Adaptability\(^22\) |
| Relevance                     | The data is sufficient for the Task and conforms to the Domain.           | Relevance\(^6\,18\,20\,23\), Relevance: Value\(^22\), Relevancy\(^10\) |
| CurrencyMeasure               |                                                                           |                                        |
| RepresentationCurrent         | Calculation for time difference between when an observation was made and when it was entered into the system. | Timeliness: delay between a change of the real-world state and...the information system\(^3\), Currency\(^1\), Completeness\(^1\), Time Completeness\(^1\), Up-datedness\(^1\), Recency\(^2\) |
| DatasetCurrent                | Time difference between when a Dataset was updated and when it was made available. For example, periodic updates to a repository. | Timeliness: ...availability of output is on time\(^3\), Opportunity\(^19\), Periodicity\(^18\), Currency\(^1\), Timeliness: Data currency\(^23\), Timeliness\(^11\) |
| TaskCurrency                  | The Data is sufficiently up-to-date for the requirements of the Task.      | Timeliness: ...information is up to date for task\(^11\), Timeliness: ...age of the data is appropriate for the task\(^11\), Timeliness (external)\(^10\) |

Table 2: Data Quality Ontology - Measure Detail
**Illustrative Example of Using the DQ Ontology**

In what follows, an example is provided to illustrate the utility of the DQ ontology concepts. Table 3 lists constraints (using pseudo-code) for some of the **Measures**. These will be used to show how data quality measures can be computed for a sample **Dataset** (Table 4) with respect to the task of calculating an eMeasure. An eMeasure is a ratio for a health outcome of interest. For example, NQF 0018, “Controlling High Blood Pressure”, is defined to be “The percentage of patients 18-85 years of age who had a diagnosis of hypertension and whose blood pressure was adequately controlled (<140/90mmHg) during the measurement period.”

<table>
<thead>
<tr>
<th>Measure</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>RepresentationConsistency</td>
<td>Representation is valid format</td>
</tr>
<tr>
<td>DomainConsistency</td>
<td>RepresentationConsistency and (Representation DomainConcepts are in Domain) and DomainComplete and Representation’s DomainConcept Constraints are satisfied</td>
</tr>
<tr>
<td>CodingConsistency</td>
<td>if Representation is coded text then Representation should have valid code</td>
</tr>
<tr>
<td>DomainMetadata</td>
<td>Domain ontology is consistent</td>
</tr>
<tr>
<td>RepresentationComplete</td>
<td>Representation value is not empty</td>
</tr>
<tr>
<td>DomainComplete</td>
<td>DomainComplete or Representation’s DomainConcept cardinality is satisfied</td>
</tr>
<tr>
<td>Sufficiency</td>
<td>Task SufficiencyConstraint is satisfied</td>
</tr>
<tr>
<td>DomainCoverage</td>
<td>Domain’s DomainConcepts are subset of Dataset’s DomainConcepts</td>
</tr>
<tr>
<td>TaskCoverage</td>
<td>DomainCoverage and (Task’s DomainConcepts are subset of Dataset’s DomainConcepts)</td>
</tr>
</tbody>
</table>

Table 3: Examples of Data Quality Measure Constraints

For the DQ ontology to be applicable, a **Domain** and a **Task** need to be defined. In this case, the **Task** is to calculate the eMeasure defined above and the **Domain** consists of concepts related to blood pressure as well as some information about the patient and the encounter. To make the example more concrete, a minimalist (and incomplete) **Domain** and **Task** ontology will be defined. A portion of a blood pressure (**Domain**) ontology is shown below (patterned after the openEHR blood pressure clinical model):

**BloodPressureDomain** (portion) is an instance of a **Domain** ontology consisting of:
- Patient is a Structure and has 1 MRN, [0 or more] Encounter, 1 Age
  - Age is a Quantity with a constraint of “Age > 0 and < 120”
- Encounter is a Structure with [0 or more] Diagnosis, [0 or more] BloodPressureObservation
- BloodPressureObservation has [0 or 1] Systolic, [0 or 1] Diastolic
- Systolic is a Quantity with a constraint of “value > 0 and < 1000, Systolic > Diastolic”
- Diastolic is a Quantity with a constraint of “value > 0 and < 1000, Systolic > Diastolic”

The **Task** usually has a formal ontology, but for simplicity’s sake a task definition serves to illustrate how concepts in the **Domain** are referenced to specify the criteria for the patient population of interest. It defines the semantics of “diagnosis of hypertension” which, in this example, is a value set of codes from the ICD9 terminology. A portion of an example **Task** instance, TaskNQF0018 is shown below. It is patterned after the eMeasure Quality Data Model (QDM):

**TaskNQF0018** (portion) is an instance of a **Task** ontology consisting of:
- PatientPopulation refers to Patients Age and Diagnosis:
  - InclusionCriteria: Diagnosis in \{401.0, 401.1, 401.9\} and Age \(\geq 18\) and Age \(\leq 85\)
  - SufficiencyConstraint: At least 1 BloodPressureObservation per Encounter
  - Numerator refers to the most recent BloodPressureObservation: Formula is count( BloodPressureObservation.Systolic > 140 and BloodPressureObservation.Diastolic > 90 )
  - Denominator refers to PatientPopulation: Formula is count( PatientPopulation )

Sample patient data is shown in Table 4. Each of the cells in the table shows the value of an instance of a **Representation**. The topmost column headers indicate the **DomainConcept** to which each of the cells map. The lower column headers show the **DataValueType** for the cells in the column. For brevity, other **Representation** information (entryTime, observedTime, etc.) is not shown.
Table 4: Example Patient Data

To assess the quality of the sample data, Measurements that quantify some of the Measures were performed. For this example, the MeasurementMethod evaluates the class constraint of a Measure for all of the Representations in a Dataset and produces a MeasurementResult, which is the proportion of constraints that were satisfied. These results are shown in Table 5. The quantity in the table cell is a fraction where the numerator is the number of constraints that are satisfied and the denominator is the number of Representations for each concept. The cell also shows the decimal equivalent for the fraction. As an example, to compute RepresentationConsistency for the Diastolic DomainConcept, the three Representations in the last column of Table 4 are examined. It can be seen that these Representations have a DataValueType of numeric. But the value for Patient2 is not valid. Therefore, only two of the three Representations have RepresentationConsistency. The rest of the MeasurementResults are shown in the table.

<table>
<thead>
<tr>
<th>Domain Concept</th>
<th>MRN</th>
<th>Age</th>
<th>Diagnosis</th>
<th>BloodPressureObservation</th>
<th>MeasurementResult</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>numeric</td>
<td>numeric</td>
<td>coded text</td>
<td>numeric</td>
<td>numeric</td>
</tr>
<tr>
<td>Data Value_type</td>
<td>1</td>
<td>72</td>
<td>“ICD9:401.0”</td>
<td>147</td>
<td>92</td>
</tr>
<tr>
<td>Data Value</td>
<td>2</td>
<td>81</td>
<td>“ICD9:401.0”</td>
<td>142</td>
<td>“High”</td>
</tr>
<tr>
<td>Data Value</td>
<td>3</td>
<td>77</td>
<td>“ICD9:401.1”</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>Data Value</td>
<td>4</td>
<td>60</td>
<td>“ICD9:xxx”</td>
<td>92</td>
<td>100</td>
</tr>
<tr>
<td>Data Value</td>
<td>5</td>
<td>44</td>
<td>“ICD9:401.9”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Measurement Process Summary for Some Measures

1943
This example shows how the DQ ontology enables a meaningful discussion of data quality characteristics required for computing an eMeasure. It also illustrates a method for quantifying each Measure by evaluating the proportion of constraints satisfied by the Representations.

Discussion

The DQ ontology presented in this study harmonized data quality concepts from the literature and provides a practical framework to evaluate data quality in health care through explicit definitions using constraints and relationships between concepts. The ontological approach provides more precise definitions of concepts than simply relying on natural language, it enables computation of a quantity for a Measure (MeasurementResult) and it makes explicit the relationship between the DQ ontology and the Task and Domain ontologies. This allows the DQ ontology to be reused for different Domains and for different Tasks without having to devise new Measures. The benefit of specifying these as separate ontologies was demonstrated in the previous section. For example, when calculating the DomainConsistency Measure, constraints from the Domain ontology (i.e. “Systolic > Diastolic”) can be referenced when computing MeasurementResults without having to change the definition of the MeasurementMethod (or the computer program that implements it). The same benefit is true when calculating the Sufficiency Measure. A SufficiencyConstraint can be evaluated for different Task ontologies to yield a MeasurementResult without having to change how Measures are defined. Not having to invent a new data quality framework for every research project should make validating data quality more common and reproducible.

Precisely defining both the Domain and Task ontology are very important in accurately describing what each data quality Measure means. Some of the Measures have constraints that reference the Task; these are clearly context dependent. Other Measures reference only the Representation or the Domain and are task independent. The constraints make clear exactly how aspects of each are related and help sharpen definitions. An example will illustrate this. DomainConsistency and RepresentationConsistency often get intertwined in definitions found in the literature. Liaw6 listed a number of sub-meanings under his “Consistency” dimension. One sub-definition (“Consistency: Representation of data values is same in all cases”) is equivalent to RepresentationConsistency, but he did not list an exact equivalent to the concept of DomainConsistency. The closest mapping is “Accuracy: Refers to values and representation of output data”. On the other hand, Weiskopf32 separated and clearly defined these differences. The concept of RepresentationConsistency is embodied as “Concordance: Is there agreement between elements in the EHR, or between the EHR and another data source?” and the concept of DomainConsistency is well defined as “Plausibility: Does an element in the EHR makes sense in light of other knowledge about what that element is measuring?” But there is an issue in the “Concordance” definition in that the last part of her definition “…or between the EHR and another data source” includes reference to another Measure (RelativeCorrectness). A Representation can have RepresentationConsistency without having DomainConsistency, but the reverse is not true. This is reflected in the constraint for DomainConsistency by explicitly referring to RepresentationConsistency as part of the definition. This also highlights the usefulness of a shared vocabulary for data quality. It makes it possible to discuss nuances of data quality characteristics.

Another issue that occurs frequently in the literature is the term “accuracy;” there is an assumption that it is possible to know what is absolutely true about the world. For EHR data, there are no true gold standards for comparison. There are only other sets of data whose “accuracy” is unknown which can be referred to as relative gold standards.31 Comparing one dataset to another to yield a positive predictive value (PPV) and sensitivity measure are a useful way to characterize the data.32 The concept of RelativeCorrectness measures whether data is likely correct by matching a Representation to its counterpart in another Dataset. The matches are considered true positives and are divided by the number of Representations in the Dataset to yield a PPV as a CorrectnessMeasure. Similarly, RelativeCompleteness looks to see which “truths” of the world are captured in the EHR data. If a Representation is present in one Dataset and is also present in the other “relative gold standard”, then these true positives are divided by the number of Representations in the other Dataset to yield sensitivity as a measure of how complete the first Dataset is.

There are a number of limitations to the current research. Data quality concepts described in the meta-analyses were harmonized and mapped to concepts in the DQ ontology. Care was taken to map based on meaning or context of use, but since the meaning was from an interpretation of a definition (or sometimes, a single term), the mapping might not represent what the author of the meta-analyses intended. This research depended heavily on the core data quality concepts contained in the meta-analyses. The literature search may not have been exhaustive in finding all of the meta-analyses or there may be important data quality concepts that were not discussed in those papers. Since many data quality concepts are repeated amongst the papers, it is likely that the most important ones were captured. It is expected that additional data quality concepts will be added to the DQ ontology as the need for having a formal definition for the concept arises. Concepts that did not appear in at least three of the papers were not included in the
DQ ontology. This includes concepts such as objectivity, non-duplication, security and privacy. Future work is needed to incorporate these into the DQ ontology. The concept of DomainComplete is currently too simplistic. It will need to be expanded to better define types of missing data as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

The DQ ontology is applicable to structured EHR data. Additional research is needed to extend the DQ ontology to notes and other unstructured data present in EHRs. Natural language processing (NLP) techniques may be used to parse relevant DomainConcepts from the unstructured information. In that case, the DQ assessment techniques described in this paper could be used to characterize that portion of the data.

The next phase of this research is to use the DQ ontology to perform data quality Measurements on actual EHR data. A Domain ontology for a clinical area will be developed in full and mapped through Representations to EHR DataTypes. Similarly, a formal Task ontology will be created and referenced by the data quality Measures. The constraints for the DQ ontology Measures will be written in a formal language, which can then directly be used to compute MeasurementResults and Metrics for a real-world Dataset.

Conclusion

The healthcare data quality literature was mined for the important terms used to describe data quality concepts. These terms were harmonized into a DQ ontology that represents core data quality concepts. Four high-level data quality dimensions (CorrectnessMeasure, ConsistencyMeasure, CompletenessMeasure and CurrencyMeasure) categorize 19 lower level Measures. These concepts serve as an unambiguous vocabulary when discussing healthcare data quality. The class constraints precisely define concepts better than using natural language and provide a mechanism to automatically compute MeasurementResults to quantify data quality. The DQ ontology can be reused with different clinical Domain and Task ontologies to make validating data quality more common and reproducible.

References
Evaluating Consumer m-Health Services for Promoting Healthy Eating: A Randomized Field Experiment

Yi-Chin Kato-Lin, MS1, Rema Padman, PhD1, Julie Downs, PhD1, Vibhanshu Abhishek, PhD1

1Carnegie Mellon University, Pittsburgh, PA

Abstract

Mobile apps have great potential to deliver promising interventions to engage consumers and change their health-related behaviors, such as healthy eating. Currently, the interventions for promoting healthy eating are either too onerous to keep consumers engaged or too restrictive to keep consumers connected with healthcare professionals. In addition, while social media allows individuals to receive information from many sources, it is unclear how peer support interacts with professional support in the context of such interventions. This study proposes and evaluates three mobile-enabled interventions to address these challenges. We examine their effects on user engagement and food choices via a 4-month randomized field experiment. Mixed models provide strong evidence of the positive effect of image-based dietitian support and negative effects of peer support, and moderate evidence of the positive effects of mobile-based visual diary, highlighting the value of mobile apps for delivering advanced interventions to engage users and facilitate behavior change.

Introduction

The ubiquitous presence of the Internet offers a cost-effective alternative for information processing and communication. The potential of the Internet has been widely recognized and utilized in healthcare services, known as eHealth. eHealth has the potential to reach a broader population, make information processing easier by using graphics, and enable interactivity and personalization. It also allows asynchronous communication and has the potential to mimic face-to-face interactions. The impact of eHealth is further enhanced by mHealth, which refers to the use of mobile devices and technologies for health management. Smartphones enable intervention delivery, data collection, and evaluation without time and space constraints, thus increasing treatment intensity and reducing the time and resources needed for application implementation. Leveraging the benefits of both eHealth and mHealth, smartphones with Internet connectivity and built-in functions, such as cameras, make it especially easy and convenient for users to record and transmit rich information at their convenience.

As a result of the ease with which it embeds in people’s daily lives, the use of mHealth interventions to influence consumers’ lifestyles has grown immensely, particularly with the boom in health and fitness smartphone apps. Popular apps include LoseIt! and MyFitnessPal for weight management, and Calorie Counter and Fooducate for healthy eating. However, the effect of mobile-enabled behavioral interventions on health behaviors is still uncertain, as consumer health or behavioral outcomes have only been evaluated for automated SMS, which have been found to usually have little or no effects, or been evaluated for the app as a whole but not isolated components. Although few apps with advanced features have been evaluated for behavioral outcomes, the evidence is only supported by observational studies (e.g. Helander et al. (2014)) or randomized experiments with small sample sizes (e.g. Wayne and Ritvo (2014)) or short periods (e.g. Byrne et al. (2011)). This gap in practice and research thus calls for additional evidence of the effect of smartphone apps on health behaviors.

Eating habits are a fundamental behavior associated with the risk of many chronic diseases. To promote healthy eating, the United States Department of Agriculture (USDA) introduced the food pyramid in 1992, which was further updated to MyPyramid in 2005. Nevertheless, although eating habits among the U.S. population have improved over time, the dietary patterns of the majority are still far from the recommended guidelines. For example, a recent report indicates that only 23% of adults consume enough fruits and vegetables per day.

Social cognitive theory suggests that people’s health behaviors can be influenced by environmental stimuli, or interventions, and the effectiveness of interventions can be further affected by the intervention components, intensity, and delivery mode. However, the effectiveness of interventions usually cannot be attributed to isolated components, as studies often evaluate the effect of a program as a whole with mixed elements. In addition, current interventions are usually delivered via traditional approaches and their treatment intensity has usually been low. To leverage the advanced functionalities of smartphone apps, this study proposes “upgraded” interventions to be
delivered by a smartphone app and evaluates their effects on user engagement and eating behaviors, on the basis of three commonly used intervention components.

The first intervention component is **self-monitoring, or dietary tracking**, which is the building block for behavioral change. Traditionally, it has required recording the amount of each consumed food item using unfamiliar measurements such as cups, ounces, or servings, which have been found to be burdensome and difficult to sustain. In addition, consumed portion size is commonly misperceived, which may lead to the consumption of more or less food than desirable.

The second common intervention component is the **support of healthcare professionals**, which aims to improve people’s knowledge to perform the desired behavior. Before eHealth, such support was usually provided during face-to-face encounters. However, the intervention intensity was low, as the encounters usually occurred only once over the course of weeks or months. When issues arise between sessions, this infrequent support usually fails to provide timely feedback. In addition, without objective references about consumers’ food intake, dietitians’ feedback based on users’ biased food diaries cannot be optimal. Although eHealth and mHealth enable frequent communication between professionals and individuals through forums or SMS, the information exchanged still may not contain objective references to consumers’ food intake. Moreover, without receiving meal-based feedback that can help individuals adjust their perceptions, people are still challenged to translate recommendations into practice.

The third commonly used intervention is **peer support**, which connects people with the same conditions or goals so that they can relate to each other’s beliefs and feelings. Before eHealth, peer support was provided via infrequent group sessions or phone calls, which was also low in intervention dosage. The use of eHealth and mHealth connects each individual online with peers 24/7, mostly via forums or sometimes via following or tracking others’ activities. However, the effect of online social support on health behavior or outcomes is still uncertain, and the literature consists mostly of observational studies, but not randomized controlled trials. Support from remote professionals and peers are not only made more accessible by the Internet, but also strongly desired by people; yet, how peer support interacts with professional support is an important but understudied issue. Specifically, while peer support may engender no effect or unexpected effects on some occasions, it is important to know whether it combines with professional support to become highly effective, to reach a saturation level, or to interfere with one another. However, studies rarely compare these two sources of support systematically together and thus the interaction effect remains unknown.

To address these limitations, this study proposes three mobile-enabled interventions that build on the three traditionally used interventions mentioned above. First, we propose the use of smartphones’ built-in cameras and the concept of MyPlate, a visualized food guide launched by the USDA in 2011, as the basis of visualized and simplified self-monitoring. It introduces a visual plate as a general guideline at mealtime and helps consumers think about their entire meal instead of counting calories, cups, ounces, or servings. Specifically, we ask users to take a picture of their meals using our smartphone app, compare their meals with MyPlate, and rate the portion size of each food category as just right, undersized, oversized, or none. This visual diary is intended to overcome the burden of traditional self-monitoring, as well as provide a reference for peer and professional support.

Secondly, with instant access to the online database that stores users’ image-based food records, the dietitian is able to provide meal-based feedback and calibrate users’ perceptions by acknowledging the ratings as their perceptions and using the food images as objective references. In addition, the continuous food records also enable the dietitian to review users’ meals for the entire week and provide feedback on their meal patterns. Both the feedback on specific meals and weekly patterns are intended to help users to adjust their eating behavior on the basis of their actual experiences.

Lastly, peer support is provided via a private version of social media, which is available only to our study participants. The social media application is designed in such a way that users can create and join groups of interest, as well as create, like/dislike, and comment on posts that may include text, images, and ratings of portion sizes with just a few clicks. Facilitated by smartphone apps, our study is also able to evaluate the interaction effect between professional and peer support at low cost to all the stakeholders, including the researchers, dietitians, and participants.

In summary, the purpose of this study is to compare the effect on user engagement of a visual diary enabled by a smartphone app to a web-based non-visual diary that requires users to record portion sizes in cups or ounces. For the
mobile-based visual diary, we systematically explore the effects of adding more timely and personalized dietitian support and peer support, both alone and in combination. We examine the effects of these external support mechanisms on engagement in self-monitoring and food choices. The effectiveness of our interventions is evaluated via a 4-month randomized field experiment. With rigorous studies about mobile apps being sparse, this study is expected to provide scientific evidence of the effects of our two innovative mobile interventions: visual diary and professional support. In addition, facilitated by mobile apps, this study is also expected to provide new evidence on the hitherto unknown interaction effects between peer support and professional support.

**Methods**

**Experiment design**

Our intervention effects were examined using a 4-month randomized field experiment. Participants were randomly assigned to one of five intervention arms, four of which included access to different versions of a mobile app, and a fifth control arm which used a web-based non-visual diary and no interpersonal support. The four mobile-app treatments all included the visual self-monitoring interface, with extra features derived from a 2 (peer support: present or absent) x 2 (professional support, present or absent) factorial design. The experiment design is summarized in Table 1 and the descriptions of each intervention will be detailed in the Interventions section. Participant demographics, food-related characteristics, and other usage behaviors were collected via a baseline survey and four, monthly, post-intervention surveys. Behavioral data on food consumption and app and feature usage were collected by our intervention tools during the entire course of the experiment, but were summarized as monthly measures. The entire experiment was conducted online without any physical encounters. The study protocol was approved by the University’s Institutional Review Board and was registered with Clinicaltrials.gov.

<table>
<thead>
<tr>
<th>Table 1. Experiment design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Web-based non-visual diary (W)</td>
</tr>
<tr>
<td>Mobile-based visual diary (M)</td>
</tr>
<tr>
<td>Peer support (P)</td>
</tr>
<tr>
<td>Image-based dietitian support (D)</td>
</tr>
</tbody>
</table>

**Participants**

Individuals were eligible to participate in the study if they were Android smartphone users, at least 18 years of age, living in the United States, and not following specialized diets such as a diabetic diet prescribed by the participants’ doctors, fad, crash, or other diets that require drastic restrictions. 425 participants were recruited between March, 2014 and June, 2014. Following the completion of a brief screening and an online consent form on Qualtrics, participants were randomly assigned to one of the five intervention arms and were immediately notified of the assignment. Participants were then asked to complete a baseline survey and were compensated $8 with a chance to win $200 for the time spent in completing the baseline survey and all four monthly surveys. Access to the intervention tool (an Android app or a website) was then provided via email, including a username and password.

**Interventions**

Participants in all five arms received an education package that highlighted the importance of healthy eating, introduced the idea of MyPlate, and provided a tailored daily food plan suggested by MyPlate, downloaded from ChooseMyPlate.gov (http://www.choosemyplate.gov/supertracker-tools/daily-food-plans.html). They also received a video tutorial for the assigned intervention tool, a prompt for goal setting, and two reminders via email per week. All participants in the four mobile-app arms received the smartphone app with features corresponding to their randomly assigned condition. The control arm participants received the uniform resource locator (URL) to the web-based application.

**Mobile-based visual diary (M):** The basic function of the mobile app was to allow participants to record their food consumption by taking up to four pictures of their food per meal, providing text descriptions, and rating the portion sizes. Ratings were assigned by the participants according to their perception about whether the consumed food group is just right, undersized, or oversized according to MyPlate’s guideline. As can be seen in Figure 1a, after
taking a picture of the meal, the user describes the meal in the text box, rates the portion size of fruits, vegetables, protein foods, grains, and dairies by clicking the colored boxes, and chooses the consumed portion size. The plate image changes according to the selected portion sizes. They also check whether the meal has saturated fat, high sodium, and added sugar in the bottom of the screen, although this information was not used in this study. After the meal entries are submitted, the application compares the participants’ accumulated food consumption for the day with their daily goals, and provides suggestions for the rest of the day to help participants achieve their daily goal. Participants can also find statistical reports on their historical goal attainment.

Peer support (P): A platform was provided as part of the smartphone application, allowing those participants randomized to intervention arms with peer support to join in social groups in the platform. Separate platforms were maintained for each of the two arms, with and without dietitian support, to prevent cross-group contamination. Once they joined a social group, participants were also able to read, like, dislike and reply to existing posts, or to post new texts/images to share in one or more social groups. The posts could include text, food images, and/or meal ratings at participants’ discretion (Figure 1b–d). For example, Figure 1d shows a post which includes a meal picture, self-evaluations of portion sizes (the pie chart) and food components (the check boxes). A weekly digest summarizing posts in each social group was sent to all members of that group via email.

Image-based dietitian support (D): A registered dietitian was hired to provide meal-based feedback and consultation services via the mobile app to participants in intervention arms with dietitian support. Using a web-based dashboard, the dietitian was able to view meal-specific information, including time-stamped food images, descriptions, and participants’ ratings. The dietitian could also access subject-specific information, including gender, age group, daily activity level, diet pattern (e.g. vegan diet), and meal pattern (e.g. 3 meals a day). Given the meal-specific and subject-specific information, the dietitian was able to provide tailored and meal-based feedback to the participants on specific meals and on weekly meal patterns. The feedback on specific meals was provided on one randomly selected meal each week for each subject. The feedback included the dietitian’s evaluation of the portion size of the entire meal as well as Fruit and Vegetables (FV) consumption, calibration of the participants’ self-reported ratings, and suggestions to meet the recommended daily food plans (Figure 1e). The feedback on weekly patterns was provided at the beginning of each week by reviewing all food entries recorded by the subject during the previous week. Feedback included the dietitian’s evaluation of compliance with the key messages advocated by MyPlate, such as “eat colorful whole foods” and “vary your veggies” (Figure 1f). In addition to the meal-based feedback, participants also received the dietitian’s responses to their questions regarding healthy eating via the app’s message function.

Web-based non-visual diary (W): A non-visual diary that requires participants to provide an estimation of the food amount consumed in cups and ounces was published online to be used by participants assigned to the control group. The form is static and follows the format of MyPlate’s paper-based daily worksheet, which allows only text input. Participants were asked to report all meals consumed on a daily basis by classifying each food item into its corresponding food group and adding up the amount of each food group in cups or ounces each day. No reports or suggestions were provided.
Subject Characteristics: Demographic variables included gender, age group, daily activity level, race, and self-reported height and weight.

Outcomes: Healthy eating behaviors for each subject were captured by his/her engagement in self-monitoring and food choices at meal time. The intervention tools captured information on both behaviors, either through the mobile or web app. Adherence to self-monitoring was measured by the number of days that have at least one record in 4 weeks (DR). The healthiness of randomly selected meals was measured by two ratings judged by the dietician. The first was the dietician’s evaluation of the portion size of the entire meal (PS), ranging from 1 to 4. Based on MyPlate’s recommended energy consumption for each subject, a meal receives a score of 4 if the meal size is about right for the subject to meet his/her daily recommendation, and receives a score of 1 if the meal is far too small or large for that individual. This evaluation was based on a “meal-calorie” schedule that was developed by our dietician. The schedule allocated an individual’s daily calorie need to each of the meals that were usually consumed by the individual in a day, and these daily meal patterns for each individual were elicited at the baseline survey. The second rating was the dietician’s evaluation of the proportion of fruit and vegetables relative to the entire meal (FV), ranging from 1 to 4. Based on MyPlate’s guidelines, a meal is assigned a score of 4 if at least half of it is filled by fruit and vegetables, and a score of 3, 2, or 1 if it is near, far from, or very far from half of the plate. When evaluating the FV and PS scores for meals, the dietician was provided meal images, meal descriptions, and the suggested calorie goal for the subject, but was blind to the participants and their assigned intervention arms. A confidence level was also assigned to each evaluated meal indicating how confident the dietician was about her ratings, ranging from 1 (not at all confident) to 4 (very confident). Meals receiving a confidence level of less than 4 were excluded from analysis. For each subject, one meal was randomly selected for evaluation each week, but multiple ratings in 4 weeks were averaged to represent the eating behavior of the month. Therefore, the trajectories of each individual’s healthy eating behaviors in our 4-month study were portrayed by DR, FV, and PS, each having at most four data points.

Analyses
Comparisons between intervention arms at baseline were conducted using chi-square tests for categorical variables, and one-way ANOVA for continuous and normally distributed variables.

Four separate mixed models were used to evaluate the effects of our interventions on outcomes by controlling for subject random effects and fixed effects, including gender, age, daily activity level, race, Body Mass Index (BMI), and time in months of the observation since participation. The first model examines the impact of the mobile-based visual diary on DR. The remaining three models examine the impact of dietician and peer support on all three outcome variables (DR, PS, and FV). To examine the interaction effect between peer and dietician support, an interaction term between these two interventions was added to the last three mixed models. We used a normal distribution to model the distribution of the average FV and PS scores and log-normal distribution to model the engagement outcome, DR.
Statistical significance is defined at $p<.05$. Statistical analyses were performed using SAS 9.4 (SAS Institute Inc, Cary, NC), with the FREQ procedure for the chi-square test, GLM for ANOVA, and the MIXED and GLIMMIX procedures for the mixed models.

**Results**

From March to June in 2014, a total of 425 participants signed up for the study and were randomly assigned to our intervention arms. With a similar drop-out rate across arms, about 70–80 participants per arm stayed in the study for four months, which gives us a total of 375 participants. Among these 375 participants, only 239 (64%) recorded at least one meal during the entire course of the experiment and were thus defined as “users,” and only the data generated by these users were used for analyses. Table 2 shows the distributions of user characteristics for each intervention arm at baseline. The majority of the users were female, under age 25, white, sedentary, and at the boundary of normal weight and overweight. No statistical differences among intervention arms were found for any of the characteristics.

<table>
<thead>
<tr>
<th>Table 2. User characteristics at baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Gender (p=.60), n (%)</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Age (p=.86), n (%)</td>
</tr>
<tr>
<td>18-25</td>
</tr>
<tr>
<td>26-70</td>
</tr>
<tr>
<td>Activity (p=.91), n (%)</td>
</tr>
<tr>
<td>&lt; 30 min (sedentary)</td>
</tr>
<tr>
<td>&gt;= 30 min (active)</td>
</tr>
<tr>
<td>Race (p=.90), n (%)</td>
</tr>
<tr>
<td>White or Caucasian</td>
</tr>
<tr>
<td>Asian</td>
</tr>
<tr>
<td>Black, Hispanic, and Other</td>
</tr>
<tr>
<td>BMI (p=.85), mean</td>
</tr>
</tbody>
</table>

Note: M, P, D, W stand for mobile-based visual diary, peer support, image-based dietitian support, and web-based non-visual diary, respectively.

Overall, we did not find a significant effect of mobile-based visual diary on number of days of recording (DR), although the estimated median of DR for the mobile group was 18% higher than that for the web group, controlling for subject fixed effects (exp(β)=1.18, p=.80). However, there was strong evidence for the impact of the image-based dietitian support on DR. The estimated median DR for users with dietitian support was 3.74 times higher than that for users without dietitian support, controlling for subject fixed effects (exp(β)=3.74, p<.01).

As for users’ eating behavior, our image-based dietitian support did not have an overall effect on users’ FV consumption (β=.14, p=.19). Yet, the significant coefficient of the interaction term D*BMI reveals the heterogeneous dietitian effect for users with different levels of BMI (β=.04, p<.01). Specifically, without dietitian support, BMI was significantly and negatively associated with FV scores (β=-.03, p<.01), but this negative association was eliminated by dietitian support (Figure 2).

Peer support also did not have an overall effect on number of days of recording, but the significantly negative coefficient of the interaction term P*#Months shows that presence of peer support facilitated user disengagement (exp(β)=-.66, p<.01). This pattern can easily be seen by plotting the raw data for users without dietitian support, as shown in Figure 3, and a similar pattern can also be found for users receiving dietitian support. By including an interaction term between peer support and dietitian support in the models, we found that there may be “bonus effects” of having both types of support, although not statistically significant (exp(β)=2.05, p=.40 for DR; β=.01, p=.98 for FV; β=-.03, p=.86 for PS).
Discussion

Mobile-based visual diary

For the entire sample, our mobile-based visual diary did not improve users’ engagement in dietary tracking, compared to the web-based non-visual diary. The lack of a positive effect of the mobile-based diary on self-monitoring in our study is consistent with the results in some prior studies\(^{34, 35}\). Two possible reasons may account for the unexpected result. The first might be the lack of user-friendly and aesthetically pleasing interfaces, which are deemed important in models of user engagement\(^6\). The second possible reason for the result is that while some users might prefer the use of visual heuristics for convenience, others might prefer counting cups and ounces for precision. From our user survey, we did find such a split among our participants. As a result, the initial high intention brought by mobility might be impaired by the imprecise approach associated with self-monitoring. In fact, we also found from our surveys that there were two distinct groups of users with opposite preferences. Whereas one group preferred the use of heuristics, the other preferred precision. Therefore, a predictive model to suggest the best approach for self-monitoring given a person’s demographic characteristics may be an interesting line of research in the future.

Image-based Dietitian support

Dietitian support had an overwhelmingly positive impact on user engagement with strong evidence. This finding is in line with the findings in other empirical studies. For example, professional support is suggested to be effective in increasing user engagement\(^7, 36\) and diary submission\(^9\). While user or patient engagement is deemed critical in patient-centered care\(^6\), it is encouraging to know that it can be greatly improved by connecting consumers with healthcare professionals and by making professionals’ feedback really personalized and contextualized.

In addition to user engagement, our image-based dietitian support was also helpful for people with higher BMI to improve their fruit and vegetable consumption. While people with higher BMI tend to eat fewer fruits and vegetables than people within a normal weight range\(^6\), our finding suggests that image-based dietitian support is especially helpful for people who really need to improve their eating behaviors. On the other hand, while people with lower BMI might have been consuming enough fruit and vegetables, there might not be room for improvement even though dietitian support is indeed effective.

Peer support

While users’ engagement in meal recording significantly declined overtime, we found strong evidence that peer support facilitated this disengagement process. This unexpected effect of peer support might be explained by the “defensive avoidance hypothesis”\(^{37}\). The hypothesis posits that to cope with potential threats, people tend to reduce
their fear by simply not thinking about their high threat level"). In the context of our study, users might feel threatened if they believe that they are eating healthfully, but realize from social media that others are actually eating much healthier meals than them. In this case, according to the defensive avoidance hypothesis, they might avoid using the diary at all. This hypothesis might be exemplified by one incident that occurred during the study. One subject emailed the first author 2 weeks after his/her participation asking to drop out of the study just because of feelings of shame from watching other people eating so healthfully. Our analysis suggests that peer support may not work for everyone. Therefore, it may also be interesting to predict who may benefit more from receiving external support given a person’s demographic and other related characteristics.

Interaction between dietitian support and peer support

Our results show that although there may be “bonus effects” when both dietitian and peer support were provided, these effects were not statistically significant. These results suggest that at least the support from different sources do not contradict with each other, both for engagement and eating behaviors. However, under conditions of limited resources, it would be beneficial if the appropriate external support could be provided to each group of people.

Limitations and future work

This study has a few limitations. First, the evaluation of the healthiness of the meals was limited by the quality of the meal images and our assumption about the adequate portion size for the meal. Images with low quality usually resulted from a dim environment surrounding the meal setting and ambiguity about the mixture of the food items, such as in a soup or curry. Although we had excluded these images from analysis to avoid measurement errors, we were unable to avoid estimation bias as these low-quality images did not occur at random at the meal level. However, since the measures we used for analysis were smoothed over weekly time periods, we believe that the measures still fairly represent users’ average eating behaviors. In addition, we believe that these low-quality images can occur in all four mobile arms with equal probability, so the differences in the scores between intervention arms accurately represent the effect of the interventions. To evaluate users’ portion control for each meal, we used the users’ self-reported meal patterns in the baseline survey as the parameter for the “meal-calorie” schedule. However, the actual meal patterns may vary across days and thus introduce measurement errors for the portion size scores. Nonetheless, we believe that these measurement errors should also occur randomly across different intervention arms; thus, our estimates of intervention effects should still be valid. Second, due to the different user characteristics of different smartphone platforms, we are only able to generalize our results to Android users, but not iPhone users; people with higher ages are under-represented, because users with age between 18–25 account for 51% of our study sample. Lastly, because our study was conducted completely online, there were no in-person contacts with our participants. Therefore, some demographics such as BMI were self-reported and may not represent their true values.

Future studies will include analysis of the best interventions for given subpopulations. In addition, it will be interesting to verify our premise that the defensive avoidance hypothesis can explain the lack of beneficial peer effects. The networks, activity logs, and text contents in the social groups will also be analyzed to better understand how social media is used and how this usage behavior affects people’s health behaviors.

Conclusion

Drawing on Social Cognitive Theory, this study proposed and evaluated three mobile-based interventions, both in isolation and combination, in the context of healthy eating. We highlighted the value of the use of smartphones for advanced behavioral intervention delivery and efficient data collection. In particular, the use of smartphones makes photographic dietary tracking convenient, which then provides the opportunity for the dietitian to deliver contextualized feedback that greatly improves user engagement and eating behaviors. It also brings consumers closer to healthcare professionals by intensifying their communication. Furthermore, the use of smartphones for data collection also makes it easy to analyze the interactions between dietitian support and peer support.

To the best of our knowledge, this is one of the first studies in mHealth literature that evaluates the effects of mobile-enabled, abbreviated means of dietary tracking, image-based dietitian support on behavioral outcomes, using a rigorously defined and executed randomized field study over multiple time periods with a large sample size. Overall, we contribute to the mHealth literature by offering scientific evidence of the effect of three mobile-enabled intervention components. This study provides strategic insights about the design of mHealth behavioral interventions and is, we hope, generalizable to other health behaviors.
Acknowledgements

We gratefully acknowledge the use of the research funds provided by three Carnegie Mellon University faculty members, Carnegie Mellon University Graduate Small project Help (GuSH) Research Grant, and Berkman Award. Many thanks to P. Sandberg, K. Bhattacharya, B. Amoedo, and F. Battle (PHRQL Inc.) for providing access to their smartphone platform and technical support for the application, and P. Martin, Registered Dietitian at Carnegie Mellon University for providing feedback on the design of the study. We are also grateful to 17 graduate students who participated in the beta testing and pilot study.

The study protocol was approved by the Carnegie Mellon University (CMU) Institutional Review Board (HS13-080/HS14-249) and was registered with ClinicalTrials.gov (NCT02206893).

References


Address for Correspondence:
Yi-Chin Kato-Lin, Heinz College, Carnegie Mellon University, Pittsburgh, PA; Email: yichinl@andrew.cmu.edu

1956
Making background work visible: opportunities to address patient information needs in the hospital

Logan Kendall1, Sonali R. Mishra, Ari Pollack, MD1,3, Barry Aaronson, MD1,2, Wanda Pratt, PhD1

1University of Washington, Seattle, WA; 2Virginia Mason Medical Center, Seattle, WA, 3Seattle Children’s Hospital, Seattle, WA

Abstract

Despite growing use of patient-facing technologies such as patient portals to address information needs for outpatients, we understand little about how patients manage information and use information technologies in an inpatient context. Based on hospital observations and responses to an online questionnaire from previously hospitalized patients and caregivers, we describe information workspace that patients have available to them in the hospital and the information items that patients and caregivers rate as important and difficult to access or manage while hospitalized. We found that patients and caregivers desired information—such as the plan of care and the schedule of activities—that is difficult to access as needed in a hospital setting. Within this study, we describe the various tools and approaches that patients and caregivers use to help monitor their care as well as illuminate gaps in information needs not typically captured by traditional patient portals.

Introduction

Hospitals are complex, dynamic, and information rich environments, yet patients and family members experience many information challenges when they are in this environment. Most hospitals have large, diverse care teams that must coordinate rapidly changing health data among providers. In addition, the information about the patient changes fairly quickly and the source of information is fragmented across hospitalists, nurses, surgeons, and different specialist services. When providing that information to patients, clinicians convey much of that information verbally and often quite rapidly. Yet, research shows that patients forget almost immediately 40-80% of what was said during a medical consultation1. Further compounding the issue, patients in the hospital are often severely ill, anxious, and stressed, which further decreases their ability to receive and process that information.

Although we know that providing outpatients with electronic information about their care can significantly improve their satisfaction with their care experience2, we know little of patient and family information needs in an inpatient context. Moreover, new initiatives to expand the role of patient portals and the pervasive presence of smart phones, tablets, and other electronic devices suggest a future where patients and families will be able to leverage real-time access to their care information while in the hospital. To explore this information space, we conducted in-hospital observations of patient-provider communication and used an online questionnaire to collect details about the experiences of patients and families. We provide a summary of the difficulty that patients and their caregivers face in obtaining and managing information while they are in the hospital and highlight information needs that are not typically a part of ambulatory patient portals. Our analysis illuminates opportunities to transform the hospital environment into an information workspace that supports greater patient and family engagement.

Background

The hospital environment creates unique challenges for patients and caregivers who are trying to access, manage and understand information about their care. We review related work that either examined the needs of patients and families in the hospital or analyzed the provision of patient-centered electronic access to medical record data.

Prior research has shown that patients perform extensive work to manage their health in the clinic3 as well as throughout their everyday lives4-6. In particular, the research suggests that patients and families are often engaged in background work, where “the workers themselves are quite visible, yet the work they perform is invisible or relegated to a background of expectation”7. In an inpatient context, patients and caregivers engage in activities to understand, organize, and manage information related to their care that is poorly understand and less visible to their care teams. Although new technologies exist to help outpatients with their work, within the hospital, patient access to information about their care is traditionally limited to verbal dialogue and occasionally to standardized, printed materials. The predominant non-verbal tool used to communicate with patients is the whiteboard, which is often used to provide information such as nurse and provider names, family contact information, and occasionally discharge expectations. Nonetheless, Marilyn Tan and colleagues found that whiteboards improved patient awareness of their care team and
the goals for their care. However, differences in use between different care providers and the difficulty in keeping the information up-to-date has constrained the effectiveness of whiteboards. Moreover, whiteboards are tools primarily managed by nurses and their design reflects a provider-managed view of determining what patients should know.

Skeels and Tan posited that technology could help patients learn about their health and care while filling the significant amount of idle time that they experience during their stay. Through inpatient interviews, the authors found that patients desired greater awareness of what was currently going on in relation to their care and more powerful and nuanced ways of communicating with their care team, families, and wider social network. In a pilot study by Weiland, et al., the authors demonstrated increased patient satisfaction and involvement in their care when cystic fibrosis patients were given a personalized schedule of their care. Simply having enhanced information about the process of care delivery in the hospital improves the patient experience. A systematic review from Prey, et al. further highlights the value of technology to support patient engagement in an inpatient setting, but also acknowledges that this space is still in its infancy. Prey and colleagues also conducted a field study where hospitalized patients were given printed copies of their raw medical record data—lab results, physician progress and consult notes, radiology reports, medication administration records—and found that even if patients did not understand all of the technical terms, they felt more informed and were able to “fact check” the information. Patient-facing technologies in the inpatient environment can improve patient awareness and management of information about their care activities.

In an ambulatory and outpatient context, health systems increasingly give patients access to their medical information through patient portals. Although evidence of their impact is still lacking, patient portals facilitate information exchange between patients and their providers and represent a larger movement to support patients in being more active participants in their care. The Open Notes project illustrates an enhanced approach to the traditional patient portal. Instead of providing a summary view of patient data and notes, Open Notes allows patients to have immediate access to the raw clinical notes created by providers as well as unfiltered access to imaging and laboratory data. In a pilot study of 19,000 patients across three institutions, most patients took time to log in and read their notes, and 99% indicated that they wanted to continue having access to their notes online. Increasingly, these types of tools show promise for supporting patients to identify medical record errors and thereby mitigate possible safety risk, but their evaluation has centered primarily in environments outside the hospital.

Other researchers have sought to enhance access to information for hospitalized patients by migrating the patient portal concept to the inpatient environment. Although limited to the emergency room environment, Wilcox and colleagues proposed a patient-facing, electronic, in-room display that would facilitate within-visit information sharing, encourage post-visit sharing and archiving, and serve as a useful memory aid as well as reference of the care delivered. They further explored the value of providing information from the medical record to give patients insight into the background activities performed by health care staff that is normally invisible to the patient. The same research group used the findings from this study to pilot a mobile-based process of care summary. Researchers have also explored a tablet computer interface to either provide access to patient portal information or deliver educational content. They found that patient awareness of their medication administration helped the patients assess their overall progress and health status. Overall, this early work suggests that creating tools to provide automated summary information for patients is feasible and highly desired by patients and families.

As patient portals become pervasive, patients are likely to have information about their health care available electronically, even in the hospital. These related studies make a case for the role that technology can play in sharing care activities (1) to promote useful awareness among patients, (2) to support decision-making, and (3) to improve patient-provider communication. However, we do not know whether a traditional patient portal designed for long-term, ambulatory use will meet the needs of inpatients. To understand the unique information needs in an inpatient setting, we explore the information work currently being done by patients and families while they are in the hospital.

Methods

We used two methods to explore the information needs of hospitalized patients: (1) observations of patient-provider interaction points conducted at two hospital facilities; and (2) an online questionnaire completed by patients previously hospitalized and by family members or other caregivers that took care of a hospitalized patient. This work was approved by the authors’ institutional review board.

Observation Study

We conducted 118 hours of observations at a children’s hospital (Site 1) and an adult tertiary care hospital facility (Site 2). The observations occurred at different interaction points with patients: physician and multi-disciplinary rounding, discharge, care conferences, physical and occupational therapy sessions, as well as ad hoc exchanges among
patients and the nursing staff. The observation findings were primarily based on open field notes and sketches of the patients’ environments. Our data reflects the perspective of care providers during their shift work, including hospitalist attending physicians, residents, specialist physicians, physical therapists, and direct nursing staff. In addition, we conducted in-room observations to observe communication and information flow from the patient’s perspective. Overall, we observed communication interactions with more than 50 individual patients. The observation notes were analyzed for factors that promoted patient information access, barriers that inhibited patient interaction during a hospital stay, and types of information managed by healthcare providers, patients, and caregivers. Our research team met regularly to review observation findings to identify preliminary themes and evaluate the extent of coverage of patient-provider communication during inpatient care.

Online Questionnaire
We also created an online questionnaire that asked participants about a prior hospital experience. Participants were recruited using convenience and snowball sampling through postings on social media sites, mailing lists, and website announcements. The questionnaire included topics across three areas: (1) importance and difficulty of receiving different types of information in the hospital, (2) approaches to managing information in the hospital, and (3) attitudes about care involvement and patient-provider communication. We received completed responses from 157 individuals from across the US and Canada who identified as being either a patient or a caregiver (e.g. family member) of a patient during a past hospitalization. We analyzed structured data from the questionnaire using descriptive summaries to identify general trends and patterns. Using an open coding technique and card sorting process24, we evaluated free text responses for related concepts and themes.

Sample population
Of 157 questionnaire responses, 61% self-identified as former patients and the remaining 39% identified as a caregiver who supported a patient during his or her hospitalization. Within the caregiver group, approximately 33% indicated they were a spouse of the hospitalized patient, 21% were parents, 21% were adult children caring for a parent in the hospital, and the remaining 25% were relatives or friends. The majority of the participants were female (75%) and ranged in age from 18 to 79. Fifty-eight percent indicated they were between the ages of 40 and 59. Respondents predominantly identified as white or Caucasian (87%). In addition, the survey participants tended to be highly educated, with 58% indicating that they have achieved some level of post-graduate education.

Forty percent (n=63) of the respondents had experienced a hospitalization within 12 months from the time they submitted the questionnaire, 23% were in the hospital between 1 and 3 years ago, and another 19% within 3 to 5 years. The reasons for the hospital admission were diverse. Many described an acute injury, while others mentioned pregnancy and various surgical procedures. Moreover, 33% of the participants described a hospital stay that lasted more than 7 days, another 29% were in the hospital between 2 and 4 days, 18% for 5 to 7 days, and the remaining 19% were hospitalized for 1 day or less. The majority of caregivers (84%) visited the patient in the hospital daily.

Results
Using mixed methods of an online questionnaire and in-hospital observations, we determined that patients and caregivers performed extensive background work to manage their information needs. To frame the context in which these stakeholders work, we begin with an overview of the patients room environment, characterized as an information workspace3. We follow these observations with an exploration of the patient’s experience receiving information based on provider workflow. By understanding the current state of information dissemination, we proceed to discuss the information gaps that patients and caregivers identified in the online questionnaire. The survey findings motivate our final observations regarding patient and caregiver work in using different tools to track information about their health.

The information workspace of a patient room
The patient’s hospital room serves as the primary focal point for information transfer among patients, caregivers, and clinical care providers. Typically, the design of the hospital room focuses on aesthetics, space for visitors, accessibility needs, support for monitoring equipment to address issues like risk of falling, noise25, and general patient satisfaction26. During our observations of patient, family, and care provider interactions in this environment, we found that different surfaces and spaces within the environment served as a means to organize, manage, and communicate care information. The design of the patient room in terms of how well the space supports information transfer is especially important considering the amount of idle time patients and caregivers experience while in the hospital.
Within the hospital rooms, we observed that patients have access to a variety of displays (Figure 1). TVs have the ability to display educational content, although that content is often standardized, rather than personalized to a patient’s situation; thus, the devices are primarily used for entertainment. Whiteboards provide a pseudo-dynamic space for patients to receive personalized information. They often have a structured layout with defined content areas for provider names, diet orders, frequency of vitals, and family contact information. Occasionally, pre-printed posters were used to display structured pathways that outlined the recovery from common surgeries like knee joint replacement. Importantly, the pathway diagrams were actionable, goal-oriented displays to assist patients in understanding the steps required before they could leave the hospital. We also observed temporary pieces of information taped above a bed or near the whiteboard that displayed instructions for nurses about patient-specific issues, such as not drawing blood from a particular arm. The displays that we observed were visible to patients and provided awareness about their care team, discharge criteria, and general care activities; however, the clinical care team primarily managed them.

Although hospital rooms have a variety of displays and surfaces to convey information to patients, we also observed challenges in being able to use these surfaces from the patient perspective. Patients had access to limited flat surface space and in-room computers frequently used by nurses were often positioned away from the bedside. We observed that the surface most accessible to the patient—a mobile tray table—was often covered with food, liquid containers, and trash that accumulate over the course of their stay. As a result, printed educational materials, notebooks, and the devices were often pushed into stacks in the corners of the room. Others described challenges with information scattered throughout a room: “Most of it I scrawled it on the back of some pamphlets that came out of a box of dressings because I didn't have any paper. They did have a whiteboard in my room so I could keep track of some of it there, but some of it was embarrassing and I didn't want it available to be read by anyone who walked in the room” (10466299). The whiteboard was rarely used by patients despite having a place for patient and family input. One caregiver described trying to use a whiteboard, “only later did we find out that it was only for staff and not for our questions for the doctor” (10276700). In our observations, many whiteboards were out of date and the information displayed was limited—typically listing only the names of care providers.

From the patient and caregiver perspective, visibility and access to information within the patient room was a challenging experience overall. To some extent, the shift to electronic systems compounds this problem because the “EMR eliminates ready review of what has been administered and when, etc. We learned to try and build good relationships with staff and were so happy when we had nursing staff that were kind, informative” (12878449). Rather than review a paper medical chart at the bedside, the patient and their family had to develop a good relationship with their nurses to stay up-to-date on care activities. Ultimately, patient room design appeared to support provider information dissemination primarily and less so the patient’s ability to utilize objects and surfaces in the room as a workspace for accessing and managing information for his or her particular needs.

**Staying informed through verbal dialogue: provider-centered access to information**

Not only are information displays in patient rooms set up primarily for provider use, but the way patients engage with information about their care is often through verbal dialogue driven by staff workflows. During rounding and other patient-provider interactions, we observed different challenges that can hinder patient and families ability to learn up-to-date information about their care. With the exception of using the call button, patients and caregivers often have to wait around for a provider to stop by in order to obtain updates about their care. Even within information exchanges like rounding and bedside handoffs, we observed providers referring to patients in the third person rather than directly engaging them in the conversation. The children’s hospital study site had instituted large, multidisciplinary rounds.
where more than a dozen providers crowded at the patient’s door and bedside. This large group creates an intimidating environment for young patients and caregivers to effectively communicate with their principal care team and to be involved in their care. Relatedly, participants in the survey were split in their agreement about being involved in decision-making (Figure 2). Fifty-nine percent agreed or strongly agreed, while 35% disagreed or strongly disagreed that they were as involved as they wanted to be. Over a third (37%) of those surveyed disagreed with the statement, “I was able to stay informed about all of the activities that occurred relating to the care provided”. Patients and caregivers appear to be very interested in being involved and staying informed but still experience communication barriers that inhibit their participation.

We also observed providers employ strategies to mitigate some of the communication challenges and help to engage patients. While observing a patient being admitted at the children’s hospital site, the admitting staff member finished talking to the parent, turned to the child patient, and told him she would need to hear from him if he felt better or worse as they treat him. She told him he was a “full partner” in this. Moreover, physicians would utilize their mobile computers or print out materials to share radiology images at the bedside. During one observation, a patient did not understand the reason why he was having difficulty swallowing food. The physician brought a computer with a large screen over to the bedside and showed the patient along with the patient’s family the latest CT scan. Not only the patient, but the patient’s entire family crowded around the display and engaged in collaborative question and answer dialogue as the physician explained his interpretation of the data. Patients and families experienced successful information exchange when their providers utilized a partnership model to collaboratively explore medical record data. Based on the experiences of survey respondents and our observations, we found that effective tool use and collaborative dialogue helped to address some of the communication challenges created by provider workflows.

**Gaps in information needs: prioritization of workflow and care activity data**

Traditionally, patients access information in the hospital through verbal interactions with care providers and occasionally through standardized, printed content. We elicited a number of information needs prioritized by patients and observed patient-driven approaches to informal information exchange with peers in the hospital.

Survey respondents identified the top three most important pieces of information that they wanted to receive during their hospital stay as: (1) medications administered, (2) the expected next visit from a nurse or doctor, and (3) lab and imaging results (Figure 3). Being able to access information on past medical history and at-home medications was less important than the other items listed in the questionnaire. Overall, respondents tended to rate information about care process activities as most important. For example, information about the “expected next visit from a nurse or doctor,” was highly rated by participants and serves as a key communication point for patients to get updates. However, this information is also difficult for hospitals to provide because of unanticipated changes in physician schedules. When asked about challenges in accessing information, more than 50% of survey participants marked the expected next visit from their provider, information about their care team, and information about patients that have experienced similar health situations as somewhat difficult or very difficult to access. As one caregiver respondent described, “We spent a lot of time sitting around waiting for the doctors. Then we would go to the bathroom or to get food, and come back and they would have been there without speaking with any of us. We would have to wait another day to ask our questions or share information or observations” (10264109).
Another notable item from the questionnaire was that 66% of patients and caregivers ranked information about the experiences of similar patients as either important or somewhat important. Despite being ranked less important than other items in the questionnaire, our observations of instances where patients engaged with their peers in informal, social interactions demonstrated the potential value of this type of information. For example, we observed one patient engaging in a physical therapy (PT) session as part of her recovery from a double-knee replacement surgery. These PT sessions often took place in a dedicated gym-like space where multiple patients and physical therapists are present. Our observed patient expressed a lot of uncertainty and doubt about her ability to climb a set of stairs but her attitude shifted after watching another patient complete the activity immediately before her and found out they both had the same type of surgery on the same day. Seeing another patient with a similar condition successfully perform the exercise, the patient appeared more confident at attempting the stair climb and succeeded in completing the activity. We also observed patients exchanging information with neighboring patients while walking around their hospital floor. Although these interactions were limited and not explicitly supported by clinicians, we observed patients providing social support and occasionally sharing information about their care experiences.

The patients and caregivers in our online questionnaire prioritized activity data—actions that have happened recently like medication administration, future plans related to scheduled tests and provider visits—when asked what information was most important. Patients also obtained valuable support through social interactions. These types of information needs indicate a desire for data about the background work of care providers and other patients.

**Patient and caregiver information work**

Based on observations and questionnaire results, patients and caregivers demonstrated active work to manage information about the care provided to them. When asked about items that they might track while in the hospital (Figure 4), survey respondents indicated that preparing questions for the care team was most important to them (90%). As one caregiver explained in the questionnaire, despite having “a robust care team...making sure we had a way to capture and communication our questions was essential” (10276700). However, their perspective on the challenge of keeping track of questions was more mixed; 47% indicated that it was either somewhat or very difficult to track. Keeping track of changes in symptoms, another item rated as important, was the most challenging item across the two groups to manage (50%). Other items, such as bowel movements or visits from family and friends, were marked as
In the survey, participants listed diverse reasons for choosing to track information during their stay. The most cited reason was to help prepare and manage questions for the next visit from their health care provider. One participant prepared captured written notes, "So we wouldn't forget our questions for the doctors, and so different members of the family could share information with each other or with the doctor or nurse when they arrived. We were never clear on when [the doctors and nurses] were coming by" (10264109). Others cited monitoring care quality as a reason for tracking. For one respondent, "The main reason I kept track was because of an error during surgery. So I knew I needed to keep a log because I wouldn't remember all the events" (11740791). Another caregiver explained how, "We wrote down "in's and outs" (nutritional intake and output) ourselves and then gave to the nurse at end of shift because this significantly increased accuracy for us to track the info ourselves" (10260629). Based on a prior event or because they believed their input was the most accurate source of data, patients and caregivers would actively record and share information with their care team.

The tools and methods used by patients and caregivers to track information about their health care varied widely. Of the respondents that stated they tracked information, almost half (49%) of these respondents used written notebooks or electronic data capture—often with a mobile phone—to document their care activities (Table 1). Some participants (18%) would keep track of things in the hospital through verbal interactions with their care team. These check-ins served as a way to reinforce what information needed to be remembered. For 5% of the respondents, they described relying on their caregivers to keep track of things for them. However, a substantial number of participants relied on their memory to keep track of things (19%). During our observations, even though many patients had a smartphone or similar device available to them, few used these devices to manage information related to their care. They served primarily as entertainment or external communication devices. We noticed that many of these devices were placed on a surface out of reach from the bedside or piled under papers and other items accumulated during the stay. The children's hospital would actually provide patients with a tablet device during their stay, but even this was often pushed to corners of the room. Patients also had to cope with physical limitations because, "with IVs in your arm and/or hand, it was difficult to write or type" (10269740). Patients and caregivers experienced difficulty using different tracking tools because of the unique constraints related to their health maintenance and the physical space of the patient rooms. The usability of the various tools also affected their ability to capture information for tracking purposes and manage information for communicating with their care team.

**Figure 4.** Within the questionnaire, percentage of Caregivers (orange square) and Patients (blue circle) that evaluated tracking different types of information while in the hospital as "somewhat important" or "important", shown on the left (a) and "somewhat difficult" or "very difficult on the right (b)."
Discussion
Through our observations and questionnaire, we found that patients often experience an information-poor environment outside of conversations with their care team. Our data illuminates patient barriers in using their hospital room as an information workspace, the difficulties that they experience in obtaining information outside of provider workflows, and their desire for information not normally displayed in the EMR or through patient portals.

Patient rooms reflect a provider-centered information workspace, despite being equipped to convey information to the patient through electronic interfaces and physical wall displays like whiteboards. For whiteboards in particular, we noticed incomplete and inconsistent usage throughout our observations that further reinforced the difficulties that patients experience in obtaining up-to-date information. Despite these challenges, the whiteboards and other displays occasionally served as useful places for physicians to diagram procedures or use visuals to help explain the patient’s condition and progress. The frequent presence of patient tools like smartphones and tablets suggest an opportunity to help patients organize all of the information collected over the course of a stay. Yet, these surfaces are frequently obscured by other objects or relegated to corners of the room away from the patient.

Within this physical environment, we found that the flow of communication is typically driven by provider workflows like rounding. Patients and caregivers experience communication challenges in this context because of the difficulty in being able to prepare for a visit that can occur at any moment. These interactions are also difficult because the information exchange is almost completely verbal and often time-constrained by the providers’ rounding schedules. In some instances, we observed providers enhance their approach to patient engagement by using collaborative behavior and electronic displays to discuss medical record data alongside the patients and caregivers. Increasing the use of visual tools at the patient’s bedside has the potential to enhance patient and caregiver participation in their care.

Finally, patients and caregivers expressed a variety of information and tracking needs related to their difficulty with managing their health information in the hospital. This extra patient and caregiver work was often motivated by concerns about care quality and wanting to stay informed when dealing with the care complexity of an inpatient environment. Accordingly, patients and caregivers indicated that information about their care team as being one of the most important and challenging pieces of information to manage. Our findings about tracking needs suggest that existing patient room displays and provider-driven information dissemination do not adequately support the patients’ and caregivers’ ability to use tools in their rooms and maintain awareness about their care activities. Moreover, the informal social interactions that we observed demonstrate that patients can obtain value through experiences of others. Reimagining the information workspace of a hospitalized patient should consider these needs and support opportunities for patient-driven care provider and peer interaction.

Addressing Information Needs
Providing enhanced access to information during a hospitalization can reduce the cognitive load for patients and improve common ground between the patient and their care team. The order that information is presented, modality, perceived importance, and health context can all have important impacts on the patient’s likelihood of remembering information. Something like a printed or electronic summary of the plan of care—upcoming activities, discharge goals,
provider visit times, and other items highlighted by our questionnaire participants—creates an opportunity for patients to process activities related to their care on their own time. They can then reassess situations with their care providers to ensure their expectations match with the documented plan of care.

Hospitals are increasingly migrating their outpatient portal into the hospital environment as one approach to addressing the problem of patient access to information about their care. This is a positive first step, but our observations and questionnaire data show that a traditional portal doesn’t address most important information needs identified by participants, such as being able to know the plan of care and when to expect the next provider visit. The interest in information about care activities suggests that an inpatient-specific portal should reflect the dynamic nature of hospital care. There is an opportunity to capture the metadata contained within order sets and to explore approaches to communicating expectations for uncertain events like when the physician will be visiting the patient next. Leveraging the process of care data embedded in the electronic medical record provides an opportunity to reconsider the patient portal as a mechanism for maintaining awareness of changes and updates in a dynamic hospital environment.

Addressing Information Work
The patients and their caregivers that participated in our questionnaire developed various strategies for coping with the information communicated verbally in the hospital. In particular, even though our respondents stated that they felt comfortable with asking their care providers questions and rated keeping track of questions as the most important activity that they could do, they also stated that keeping track of changes was one of the most challenging activities during their hospitalization experiences. This relates to our observation data showing that patients face barriers to using tracking tools. In part, this is due to constraints created by the physical environment. We often observed tracking tools such as smartphone devices, notebooks, and even pieces of paper were relegated to areas of the room distant from the patient’s bedside. Even tablets, which are increasingly used to provide access to patient portals and health records, were often out of reach of the patient. The only tool consistently at the patient’s side was the call button which can alert a nurse if the patient has a question but lacks the ability to track information. Moreover, patients and caregivers in the survey expressed how they are able to track certain types of information about themselves more accurately than the care team that may not always be present in their room. Current conceptualizations of the patient portal typically do not address tracking needs. In fact, patients and caregivers are using tools when they are available and cite the lack of resources as a major barrier to being able to manage the large amount of information and activity occurring during a hospitalization.

Limitations
Our qualitative, ethnographic approach to studying the patient experience has several limitations, including that the questionnaire sample skews toward a well-educated demographic. However, even within this group, both patients and caregivers experienced challenges with obtaining access to key information about their care. Ultimately, the questionnaire is not intended to be a representative population sample, but instead provide a means to explore a broad range of patient and caregiver experiences across the US. We integrated the questionnaire findings with observation data from two independent hospital environments help to strengthen the validity of our findings.

Conclusion
The findings of our online questionnaire paired with inpatient observations show that inpatients face many information and communication challenges during their hospital stay. Analyzing the patient’s room as an information workspace in the hospital illustrated challenges and opportunities for tools and displays that can enhance patient interaction with their care information. Moreover, when we consider the patient portal functionality that is typically used in an outpatient setting as a strategy to mitigate some of these challenges, we found that inpatients need additional support within the hospital context. In particular, patients and caregivers highly desired yet found it quite challenging to obtain the patient’s plan of care for each day. In addition, most patients wanted to track key aspects of their experience. Neither of these important functionalities are available in patient portals but often exist as metadata within the patient’s medical record. Our work highlights these important needs and suggests new functionality that emerging inpatient information systems need to support. This type of new functionality has the potential to transform inpatient care by empowering these patients and caregivers with the information they need to have a satisfying care experience.

Acknowledgements
This project was supported by grant #1R01HS022894 from the Agency for Healthcare Research and Quality (AHRQ).
References

1966
Development and Preliminary Evaluation of a Prototype of a Learning Electronic Medical Record System

Andrew J. King, MS1, Gregory F. Cooper, MD PhD1,2, Harry Hochheiser, PhD1,2, Gilles Clermont, MD MSc3, Shyam Visweswaran, MD PhD1,2

1Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA; 2Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA; 3Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Electronic medical records (EMRs) are capturing increasing amounts of data per patient. For clinicians to efficiently and accurately understand a patient’s clinical state, better ways are needed to determine when and how to display EMR data. We built a prototype system that records how physicians view EMR data, which we used to train models that predict which EMR data will be relevant in a given patient. We call this approach a Learning EMR (LEMR). A physician used the prototype to review 59 intensive care unit (ICU) patient cases. We used the data-access patterns from these cases to train logistic regression models that, when evaluated, had AUROC values as high as 0.92 and that averaged 0.73, supporting that the approach is promising. A preliminary usability study identified advantages of the system and a few concerns about implementation. Overall, 3 of 4 ICU physicians were enthusiastic about features of the prototype.

Introduction

According to the American Medical Association, reducing cognitive burden is one of the top eight priorities for improving electronic medical record (EMR) usability1. One instance of high cognitive burden is information overload where an “individual’s efficiency in using information is hampered by the amount of relevant, and potentially useful, information available to them”2. The display of patient data in current EMRs is not specific to the clinical context and burdens the physician who must search through the data to compile a clinically relevant narrative of the patient.

As EMRs capture increasing amounts of data per patient, compiling a clinical narrative becomes cognitively more demanding. This information overload is particularly challenging in settings such as the intensive care unit (ICU) where large amounts of data per patient are collected. A study in a Canadian hospital estimated that the average ICU patient generates a median of 1,348 data points per day3. This equates to a new data point being added to a patient’s record almost every minute. An EMR that focuses the physician’s attention on relevant patient data could help reduce the time needed to assess the patient’s condition and improve the quality of the resulting judgments1, enabling improved decision making, reduced medical errors, and greater efficiency.

Research efforts in the display of patient data have progressed in various directions including graphical summaries4, methods to summarize and display temporal data5, and the context specific integration of data using either systems or disease6-9 based approaches. Integrated data displays aggregate information from different sources and display it in one location (e.g., a cardiovascular system view, or a diabetes management view). These displays can improve performance and efficiency of physicians. For example, Pickering et al. developed an interface called AWARE that dynamically presented highly valued data to the physician at the ICU bedside in an integrated fashion10. When evaluated, AWARE reduced time to task completion and medical error in the assessment of ICU patients with a hypothesized acute bleed10,11.

AWARE and other existing integrated systems use rules to identify which of the thousands of available data points are relevant for specific patient contexts12. Rules are usually manually constructed from disease models, ontologies, and expert opinion. Such rule-based systems have several advantages. They are likely to be clinically informative, since they are based on clinical knowledge, and they can be readily programmed and applied to patient data that are available in electronic form. However, construction of rules is tedious and time-consuming. Moreover, rules have
limited coverage of the large space of clinical conditions, and a rule-based display may not adequately portray the
context of a patient whose condition presents in an unusual way or a patient who has multiple maladies.

An alternative to a rule-based approach is to identify data that are relevant in assessing a patient’s current clinical
condition using a data-driven approach. Such an approach identifies patterns of useful patient data from data-access
activities of physicians on past patient cases, and these patterns are applied to a patient case to highlight data
relevant for that patient. Such an approach has several advantages. It does not require expert input to develop
patterns, the patterns are derived using a large set of past patient cases, and the patterns can cover a wide range of
clinical conditions. Moreover, such a system can continually update the patterns by learning from each interaction a
physician has with it and can adapt to the specific environment in which it was being used. In other words, it would
be a Learning EMR (LEMR).

Appropriate user-interface design will be vital to the success of a LEMR. Displays that adapt to suit the context of
use have the potential to optimize the use of scarce resources, including user attention and screen-display real
estate. To be successful, these designs must provide dynamic displays that inform without disorienting. Recent
efforts have explored novel designs for displays of laboratory data, configuration of multiple components in an
EMR display, and for knowledge-based adaptations of order entry menus. A successful LEMR would build
upon these results, using a combination of appropriate interface design techniques and arrangements with
context-sensitive adaptations based on patterns of system use.

We built a prototype LEMR that uses a data-driven approach to predict and highlight the data that are most relevant
to a patient case. A physician reviewed patient cases and identified the relevant data for each case. The identified
data became the dataset for training and evaluating classification models that predict if a data item is important to
understanding the current patient’s state. The results are used to determine if the data-driven approach is promising.
Four more physicians used the prototype LEMR in a usability study. We describe the physician reactions to both the
LEMR concept and the prototype. The results of the usability study provide an initial assessment of the desirability
of such a system and an evaluation of the prototype design.

Methods

1. LEMR prototype

For this study the LEMR prototype required (1) screens to display patient data, (2) a dynamic region of the screen
that could be filled with the clinically relevant data items, (3) the ability to record which data items a physician
found clinically relevant, and (4) a framework to connect all of the components.

Highlighted Information Display (HID)

The LEMR prototype needed a clear and intuitive way to indicate which items were clinically relevant for a patient
case. For this purpose, we allocated a region of the screen to be dynamically populated with the clinically relevant
items. We called this region the Highlighted Information Display (HID) because the items found in it are the items
that should stand out to the physician user.

Recording data-access patterns

We relied on physicians manually selecting the items they found most clinically useful for the most recent day of
each patient case they reviewed. To indicate an item as useful, the prototype had a means for the user to move it into
and out of the HID. Every time that a user modified the contents of the HID, the prototype recorded the changes in a
database.

Components of the prototype LEMR

The prototype LEMR consists of four main components: a user interface, a web framework, a patient database, and
a module with statistical models. These components are shown in Figure 1. The user interface is implemented using
HTML, CSS, and JavaScript, and is powered by Django, an open source Python Web framework. Patient data are
stored in a MySQL patient database, which is queried to provide data shown in the user interface. This database also
stores data viewing activities that are captured during EMR use. These data are used to train statistical models that
predict data items (e.g., laboratory results) that are typically viewed in given a clinical context. These models are
then applied to a current patient case in the database to predict the data items that are likely to be viewed and thus
should populate the HID for that case.
Before enough training data are collected, the prototype relies on users viewing real patient cases and manually populating the HID with the data items in those cases that they think are most relevant. After some manual training data are collected, models start to automatically highlight the items that they predict to be of interest. When the HID is automatically populated from the models’ predictions, the user retains the ability to manually modify its contents. Any manual modifications then feed back into the statistical models to further improve them.

2. Predictive modeling

We developed and evaluated logistic regression models to predict laboratory test results that are relevant to understanding a patient’s current condition. We used de-identified data on 58 ICU patient cases that were randomly selected from a large ICU dataset that is described below. In this early study, we focused on predicting only laboratory test results using demographics and all laboratory test results as features.

ICU patient data

The patient data come from the High DEnsity Intensive Care (HIDENIC) dataset, containing fully de-identified and HIPAA-compliant EMR data on 12,000 patients who were hospitalized in ICUs at the University of Pittsburgh Medical Center (UPMC) from July 2000 through December 2001\textsuperscript{23}. HIDENIC contains structured data including demographics, physiological measurements collected at the bedside (including vital signs), laboratory values, medication and fluid administration records, and unstructured data in the form of a variety of clinical text reports (history and physical, progress, and operative procedure notes; radiology, EKG, and EEG reports).

The HIDENIC dataset contains patient information from ICU admission until ICU discharge. Because we were interested in physician behaviors during the review of an average day of an ICU patient, we defined the data for a patient case to be from admission to a random day during the patient’s ICU hospitalization. Using this dataset, we were able to show a patient case to a physician and have them act as if the case was a current patient of theirs.

Extracting the features

The features used for predicting the clinical relevance of each laboratory test were extracted from each patient case and included both temporal and non-temporal items. Temporal features included the most recent value of each laboratory test and a Boolean value for whether that test result had appeared within the last 24 hours of available patient data. There were 190 distinct laboratory tests in the dataset, so the temporal test features totaled 380 (= 190 * the two features extracted from each test). There were five non-temporal demographic features (age, sex, weight, height, and body mass index) and one additional temporal feature of days since ICU admission. Therefore, in total there was a set of 386 predictive features for every patient case.

Collecting the training cases

Each of the 190 laboratory tests defines a binary target (dependent) variable that indicates whether that test is clinically relevant in a given patient case, and therefore should be placed in the HID. To collect this target data, 59 patient cases were randomly selected from the HIDENIC dataset. The cases were loaded into the LEMR prototype and a single physician reviewer (author SV) used the prototype to identify the tests that are relevant for each patient case within a given clinical context. In particular, the reviewer imagined that he was the attending who was taking care of the patient. He read the clinical reports and examined the test results to determine the clinical course since admission to the ICU. Then, for the last day for which patient data were displayed, he identified which tests were relevant in providing evidence about (1) changes in the clinical condition of the patient, or (2) the emergence of a new clinical problem. Any test that the reviewer moved into the HID for a patient case was considered clinically relevant for that case and was combined with that patient’s feature set to create a positive training example for that test. Any test that the reviewer did not move into the HID for a patient case was considered not clinically relevant for that case and was combined with that patient’s feature set to create a negative training example for that test.
Modeling and evaluation

Each laboratory test had a training dataset constructed from the 59 patient cases that were reviewed. We used each test’s training dataset to train a penalized logistic regression model that predicts if that test is relevant to the current patient case. The models were build using the scikit-learn Python package24 and evaluated using leave-one-out cross-fold validation. We measured performance using the Area Under the Receiver Operating Characteristic curve (AUROC).

3. Usability study

A usability study was conducted to gauge physician assessment of the LEMR concept. Four fellows in UPMC’s Department of Critical Care Medicine participated in the study. This study was approved by the University of Pittsburgh Institutional Review Board (ID PRO14020588). Each physician used the prototype to review three to five selected patient cases. For each case, the physicians were shown a patient’s data from ICU admission to a random day during that patients ICU hospitalization. The physicians were asked to familiarize themselves with the case as if they were the attending physicians. No data were highlighted in the HID for the initial look at each patient case. Next, the physicians were shown an additional day of data that was meant to simulate rounding on the subsequent day. For the additional day, items had been manually highlighted, based on relevant items identified by a physician on the research team (SV). The physicians were asked to use the features of the prototype to highlight/unhighlight items until the highlighted items represented the data that they thought another physician who was looking at the same case would want to see when assessing the most recent 24 hours, given that they had been following the patient case during the current ICU admission.

During the review of each patient case, screen tracking software recorded all of the on screen actions and an audio recording captured each physician's think-aloud comments. After a physician reviewed the allotted patient cases, additional time was allocated for a semi-structured interview of him or her. We asked about their perceptions regarding the LEMR concept in general and the prototype specifically. The interviews were coded independently by two researchers (AK and HH) before meeting to create a consensus. The coding was used to identify general themes in the responses. Each physician also filled out the System Usability Scale25 based on their interactions with the prototype.

Results

1. Prototype

A screen shot of the user interface of the LEMR prototype is shown in Figure 2. Panel A, the patient demographics toolbar, allows the user to move between patients and gives a brief summary of the current patient’s demographic information and admitting diagnosis. Panel B contains quick access tabs for navigating among the various types of patient data, including laboratory test results, medication orders, and clinical text reports (e.g., history & physical (H&P) reports, progress notes, and operative procedure notes). Currently the “Labs/Vitals/Meds” tab is selected. This prototype uses times-series plots to display these non-text based clinical items (laboratory test results, vital signs, medication orders, and intake and output data). Panel C, the time range selector, is used to define time ranges of data to display. Below the time-range selector is the procedures axis, which is labeled with the defined times. Black diamonds on this axis represent procedures (surgeries, biopsies, etc.) that the current patient has had. Hovering over a diamond gives more details on that procedure. Panel D, the HID, shows high detail plots of selected clinical items. These plots have a labeled y-axis and blue bands to indicate the normal range. Panel E displays all available results, including those found in the HID, using plots with condensed y-axes. These plots give a notion of trends over time and are arranged by group type (chemistry, lipids, cardiac, etc.). The buttons across the top of this panel list all of the different group types and can be used to jump to a specific type. For both Panel D and Panel E, different colors are used to indicate when a value is within or outside of normal range (blue = below; green = within; red = above; black = no defined normal range).

For any given patient, the items highlighted in the HID are meant to be the ones that are most clinically relevant for that patient at the current time. The HID can be populated with items by both automatic and manual means. For automatic population, the LEMR uses stored statistical models to predict the probability that each item is relevant given the current patient state. The items that have a predicted probability above a set threshold are displayed. To make manual changes to the contents of the HID, a user clicks on the arrow buttons next to each clinical item’s name. There are buttons to move items into the HID and buttons to remove items from there.
2. Predictive models

We derived penalized logistic regression models for the 21 distinct laboratory tests that were identified as clinically relevant (accessed) for at least two patient cases in the dataset. The AUROC values for those models and the number of positive (# +) training cases in their datasets are shown in Table 1. The average AUROC is 0.73. The top seven tests shown in the table have an average AUROC greater than 0.80.

3. Usability study

The physicians were generally enthusiastic about the LEMR approach. They identified advantages, such as adaptation to different specialists and the potential for time savings. Concerns included the feasibility of the implementation and the possible implications of integration into workflow. For instance, some participants worried that over-reliance on highlighted items might cause physicians to miss important details in the remainder of the record.

Three of the physicians liked the timeline approach to displaying the data but would like to see the exact value of each result without needing to hover over a data point with the mouse. Two of the physicians discussed other information resources such as the summary sheet that they currently print out for each patient before rounding. One physician discussed a desire for the EMR to include a checklist of all the things that must be checked for each patient every day.

The System Usability Scale composite score for the four users was 78.75. The scale ranges from 0 to 100 and any score above 68 is generally considered to be above average usability26.
We developed and evaluated a prototype LEMR system that uses a data-driven approach to adapt its display to highlight the data items that are most relevant in a patient case. Our preliminary results show varying performance for the predictive models that identify when laboratory tests are relevant, but one-third of them had an AUROC of 0.80 or greater, supporting that this approach is promising. Moreover, in a usability study ICU physicians were generally supportive of the LEMR approach, and identified advantages and concerns with the prototype.

With the accumulation of data in the EMRs, physicians are at increased risk of making errors due to information overload\(^{27, 28}\). The LEMR system we are developing may help to reduce this risk by identifying and highlighting the small subset of the available information\(^ {29}\) that is relevant when making medical decisions for a patient. Similar data display techniques, such as Microsoft Word’s split menu font selection list, have resulted in faster selections and higher user performance ratings\(^ {19}\). Split menus combine the convenience of frequency ordered menus with the consistency of alphabetically ordered menus. The LEMR prototype has a similar structure where the HID provides convenience and the rest of the prototype provides consistency.

The LEMR also has similarities to other tools\(^ {18, 30}\) that have explored mining interaction patterns from previous users to inform the presentation of information/choices/etc. These other tools have benefited from crowd wisdom\(^ {31}\), where

### Table 1. AUROC for patient-specific prediction of the clinical relevance of each laboratory test

<table>
<thead>
<tr>
<th>Laboratory Test</th>
<th>AUROC</th>
<th>Lower</th>
<th>Upper</th>
<th># +</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilirubin Total</td>
<td>0.92</td>
<td>0.83</td>
<td>0.97</td>
<td>5</td>
</tr>
<tr>
<td>Liver Alanine Aminotransferase</td>
<td>0.91</td>
<td>0.72</td>
<td>0.98</td>
<td>4</td>
</tr>
<tr>
<td>Liver Aspartate Transaminase</td>
<td>0.91</td>
<td>0.72</td>
<td>0.99</td>
<td>4</td>
</tr>
<tr>
<td>PTT Coagulation</td>
<td>0.84</td>
<td>0.71</td>
<td>0.92</td>
<td>9</td>
</tr>
<tr>
<td>Lactate</td>
<td>0.83</td>
<td>0.58</td>
<td>1.00</td>
<td>2</td>
</tr>
<tr>
<td>Phosphorus</td>
<td>0.82</td>
<td>0.62</td>
<td>0.94</td>
<td>11</td>
</tr>
<tr>
<td>White Blood Cell Count</td>
<td>0.80</td>
<td>0.67</td>
<td>0.91</td>
<td>8</td>
</tr>
<tr>
<td>INR Coagulation</td>
<td>0.79</td>
<td>0.63</td>
<td>0.89</td>
<td>11</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>0.77</td>
<td>0.59</td>
<td>0.89</td>
<td>37</td>
</tr>
<tr>
<td>Sodium</td>
<td>0.75</td>
<td>0.61</td>
<td>0.86</td>
<td>18</td>
</tr>
<tr>
<td>Glucose</td>
<td>0.73</td>
<td>0.55</td>
<td>0.87</td>
<td>12</td>
</tr>
<tr>
<td>Chloride</td>
<td>0.73</td>
<td>0.59</td>
<td>0.82</td>
<td>2</td>
</tr>
<tr>
<td>Blood Urea Nitrogen</td>
<td>0.73</td>
<td>0.56</td>
<td>0.85</td>
<td>22</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>0.71</td>
<td>0.54</td>
<td>0.83</td>
<td>33</td>
</tr>
<tr>
<td>Platelet Count</td>
<td>0.70</td>
<td>0.53</td>
<td>0.82</td>
<td>28</td>
</tr>
<tr>
<td>Lymphocytes Absolute Count</td>
<td>0.64</td>
<td>0.26</td>
<td>0.95</td>
<td>2</td>
</tr>
<tr>
<td>Neutrophils Absolute Count</td>
<td>0.64</td>
<td>0.27</td>
<td>0.95</td>
<td>2</td>
</tr>
<tr>
<td>Red Blood Cell Count</td>
<td>0.57</td>
<td>0.25</td>
<td>0.97</td>
<td>3</td>
</tr>
<tr>
<td>Magnesium</td>
<td>0.56</td>
<td>0.27</td>
<td>0.89</td>
<td>5</td>
</tr>
<tr>
<td>Potassium</td>
<td>0.52</td>
<td>0.37</td>
<td>0.68</td>
<td>11</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.47</td>
<td>0.28</td>
<td>0.83</td>
<td>5</td>
</tr>
<tr>
<td>Average</td>
<td>0.73</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

We developed and evaluated a prototype LEMR system that uses a data-driven approach to adapt its display to highlight the data items that are most relevant in a patient case. Our preliminary results show varying performance for the predictive models that identify when laboratory tests are relevant, but one-third of them had an AUROC of 0.80 or greater, supporting that this approach is promising. Moreover, in a usability study ICU physicians were generally supportive of the LEMR approach, and identified advantages and concerns with the prototype.

With the accumulation of data in the EMRs, physicians are at increased risk of making errors due to information overload\(^ {27, 28}\). The LEMR system we are developing may help to reduce this risk by identifying and highlighting the small subset of the available information\(^ {29}\) that is relevant when making medical decisions for a patient. Similar data display techniques, such as Microsoft Word’s split menu font selection list, have resulted in faster selections and higher user performance ratings\(^ {19}\). Split menus combine the convenience of frequency ordered menus with the consistency of alphabetically ordered menus. The LEMR prototype has a similar structure where the HID provides convenience and the rest of the prototype provides consistency.

The LEMR also has similarities to other tools\(^ {18, 30}\) that have explored mining interaction patterns from previous users to inform the presentation of information/choices/etc. These other tools have benefited from crowd wisdom\(^ {31}\), where
the collective knowledge of many people can be better than the knowledge of any individual in the population. In
terms of a LEMR, a physician who is inexperienced with a certain patient type could benefit from seeing items
highlighted based on the information that a more experienced physician would access. This would bring the
inexperienced physician’s attention to an important item that they may have otherwise overlooked. Highly
experienced physicians could perform the initial training of the system to further capitalize on this effect.

The AUROC performance of the models developed during this study ranged from ~0.50 to 0.92. The higher scores
are a promising sign for the approach because we used a small number of training cases and a limited set of possible
features. A dataset of only 59 cases leaves only 58 training cases for each fold of the validation. If a LEMR were
implemented in a hospital setting, it would learn every time a physician interacts with it. That would mean
additional training cases extracted for each patient every day. Using this approach, the size of the training set would
quickly grow to a huge number of samples. The feature set used in this study was purposely restricted to very basic
predictors. Many more temporal features, such as the slope between the last two values of a laboratory test, could
prove to be very useful when predicting if a test is relevant to a patient context. Numerous other data types, such as
vital signs and medication orders, were not considered but could be included as well. With an increase in the training
set size and an expanded feature set, the predictive accuracy of each model would likely improve.

Our usability study provided preliminary support for the feasibility of the LEMR idea. Participants’ concerns
regarding feasibility and integration into existing workflows are valid practical points that we hope to explore after
modeling and interface design issues are more mature. The possibility of over-reliance on highlighted items is a key
concern—EMR displays that encourage overly narrow readings of patient information may cause important data to
be missed. Design lessons from prior work in personalized interfaces and related areas will inform the
development of prototypes exploring alternative displays, with continued usability studies providing vital physician
feedback.

The continual extraction of training data is the biggest advantage a LEMR has over rule-based context sensitive
EMR systems. Rule-based systems do not generalize well and their display of data may not be optimal for any
patient who does not fit into a planned context. For example, a patient can have both diabetes mellitus and chronic
kidney disease. A display that was designed specifically for a patient with diabetes may not adequately represent the
complexities of the case. The LEMR, which learns from every interaction a user has with it, could represent some of
the important complexity of any case and could learn to meet information needs as they change overtime.

The data-driven approach we are taking also allows for the same LEMR system to be implemented in almost any
healthcare setting. For example, if the LEMR were implemented in a particular unit at a hospital, the combined
interactions from all of the physicians in that unit would aid the system in adapting to how patients are handled in
that environment. The system would not be limited to one set of models either; it could use hierarchical learning to
have a different set of models for each class of users, such as physicians of different specialties. The same theory
would apply when implementing the LEMR in a hospital wide setting. The models could customize to the different
groups of users in the different units.

Limitations

Our study has several limitations. The training of the predictive models was limited by the small training dataset. It
contained enough data to evaluate the prediction of 21 laboratory tests. This limitation will be addressed in a future
study where more patient cases are reviewed and other types of clinical data are included in the items whose access
is tracked. Moreover, we plan to develop predictive models for data items beyond laboratory test results.

The usability study was limited to four participants. In the future we plan to conduct usability studies with a larger
number of participants, potentially including side-by-side empirical comparisons aimed at understanding the impact
of alternative design strategies on physician interactions with the EMR, in terms of both accuracy and efficiency.

We obtained data-access information from a physician user who manually highlighted data items that were relevant
to a patient case. This manual acquisition of data-access information is a limitation of the current study. In a mature
LEMR that is deployed, such data-access information would be inferred automatically from computer mouse
activity, eye movement tracking, and from clinical notes (e.g., progress notes) in which pertinent data items are
recorded.
Conclusion

EMRs should synthesize and present patient data in such a way that it enhances the physician’s ability to understand the overall state of the patient and detect significant changes in the clinical course while decreasing the physician’s cognitive workload. We proposed a data-driven approach to displaying EMR data in a context specific manner. We built a prototype, trained and evaluated predictive models, and conducted a usability study as first steps in the development of a LEMR. This study supports the feasibility of the approach and provides insight on development concerns and opportunities. We plan to use the lessons learned from this study to extend the models using larger and more comprehensive datasets, to further refine the interface, and to conduct additional evaluation studies of the LEMR.

Acknowledgements

We thank Mr. John Levander for his help in developing the LEMR prototype. We thank Dr. Milos Hauskrecht for providing the ICU data that was extracted and processed in another project. Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award numbers T15LM007059 and R01LM012095, and by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM088224. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


22. Django. 1.5 ed. Lawrence, Kansas: Django Software Foundation; 2013.


A Multidimensional Data Warehouse for Community Health Centers

Kislaya Kunjan, MS, MBA1, Tammy Toscos, PhD1,2, Ayten Turkcan, PhD3, Brad N. Doebbeling, MD, MSc1

1Department of BioHealth Informatics, Indiana University, Indianapolis, IN; 2Parkview Research Center, Fort Wayne, IN; 3Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA

Abstract
Community health centers (CHCs) play a pivotal role in healthcare delivery to vulnerable populations, but have not yet benefited from a data warehouse that can support improvements in clinical and financial outcomes across the practice. We have developed a multidimensional clinic data warehouse (CDW) by working with 7 CHCs across the state of Indiana and integrating their operational, financial and electronic patient records to support ongoing delivery of care. We describe in detail the rationale for the project, the data architecture employed, the content of the data warehouse, along with a description of the challenges experienced and strategies used in the development of this repository that may help other researchers, managers and leaders in health informatics. The resulting multidimensional data warehouse is highly practical and is designed to provide a foundation for wide-ranging healthcare data analytics over time and across the community health research enterprise.

Introduction
Primary care practices and Community Health Centers (CHCs) are key components of the health care system in the United States (US), providing chronic and preventive care services to low income and medically underserved areas1. Incorporating health information technology (IT) systems for healthcare delivery has been recognized as a means to support improvements in patient outcomes, while yielding cost savings for patients, providers and payers2. CHCs are atypical in terms of their patient population, which in turn affects their overall clinical operations, including their health IT requirements. With over 96% of the CHCs using EHRs and handling over 90 million patient-visits a year3, they are an excellent source of clinical and administrative data and an ideal setting for research and development. However, when data resides in multiple disparate silos, the payers and providers cannot aggregate, analyze and assess the data in a cost-effective manner4. A data warehouse that can aggregate multiple disparate data sources, including patient records, clinical quality scores, and payer data for CHCs can be used to efficiently improve clinical and financial outcomes across the practice and the region at large5, 6, as well as to foster informatics and operations research.

In addition to the state-wide operational health information exchanges (HIEs) in the US,7 there exist several large data repositories/warehouses of significance, some of which include: Healthcare Cost and Utilization Project (HCUP)8; National Patient-Centered Clinical Research Network (PCORnet)9; Center for Medicare and Medicaid Claims Data10; US Food and Drug Administration’s Sentinel Initiative11; CancerLinQ12; and eMERGE (Electronic Medical Records and Genomics)13. Enterprise data warehouses to manage heterogeneous biomedical data have been developed for medical centers, such as Veteran Health Administration’s Corporate Data Warehouse14, Intermountain Healthcare’s CIS15, Massachusetts General’s COSTAR16 and Mayo Clinic’s Enterprise Data Trust17.

A CHC data warehouse that combines the best features of the HIEs and consolidated repositories like HCUP, along with health data analytics, can be particularly desirable for CHCs. In this paper we share our journey to create a multidimensional clinic data warehouse for seven clinics in the state of Indiana with whom we are collaborating to improve access to care for underserved patients. The primary research question guiding this work is how to design a CDW to support improvements in care delivery at community health centers.

Background
Indiana is home to the Indiana Health Information Exchange (IHIE)18, which is the largest HIE in the US. However, Regional Health Information Organizations (RHIOs) that manage such local HIEs have limited utility for CHCs because the participating facilities that share data into HIEs are typically hospitals and integrated health centers, in contrast to community clinics that are free standing and resource constrained. There are additional challenges with
HIEs such as: high start-up (upwards of $12 million) and operating costs; challenges in their business model, need for a dedicated infrastructure to managing streaming data from the healthcare facilities, while addressing issues of privacy and security, creating and maintaining a master patient index or record locator service and finding technologically capable exchange partners; and dealing with issues of distrust and data control issues among competitors. Furthermore these HIEs are primarily focused on patient-identifying clinical data derived from EHR and ancillary systems with a focus on allowing data to follow patients across delivery settings. They are not designed to be data stores or a data warehouse which capture operational and financial data and/or provide data analytics to members in the exchange network. In our research, we discovered that the CHCs in Indiana were not part of the IHIE member network, not much unlike the 280 other statewide operational HIEs in the nation who lack CHC involvement. Most CHCs we interviewed expressed keen interest in being able to effectively draw insights from their data, even as they encountered challenges in doing so. We chose to develop a unique community clinic data warehouse that would capture not only the medical record data but also include operational and financial data of the clinics in a very cost-efficient manner. To this multidimensional data warehouse, we intend to add an analytics platform, with a goal to support better patient care which can be defined in terms of improved access to care, shared decision making, and greater patient satisfaction. In addition, the intent is to enhance the operational and financial health of the community health care systems in the state of Indiana. We also believe there is great opportunity to foster community informatics research and development using this data to improve care delivery, patient engagement and efficiency.

The proposed work is an offshoot of an ongoing 3-year research project funded by the Patient Centered Outcomes Research Institute (PCORI). The larger project aims to improve access to care for CHCs by: identifying common barriers and successful patient-centered strategies; applying Delphi expert panel methods along with workflow and simulation modeling to identify feasible and valid patient-centered, strategies; and finally determining if implementing patient-centered systems can improve access to care, relative to existing approaches. The entire study, including the data collection from the clinics, has been IRB approved by the Indiana University Office of Research Administration.

Methods
We partnered with 7 CHCs geographically spread across the state of Indiana. The clinics were selected such that they represented urban, semi-urban and rural settings and provided care to diverse patient population in terms of race, ethnicity, insurance status, and income levels.

Design of the Community Clinic Data Warehouse
A common and widely accepted definition of a data warehouse is that it is a subject-oriented, integrated, non-volatile, and time-variant collection of data to support decision making. Data warehousing methodologies share a common set of tasks, including business requirements analysis, data design, architectural design, implementation and deployment. We present a data warehouse architecture that is simple, low cost, and one that can readily benefit resource constrained CHCs.

Our multidimensional Clinic Data Warehouse (CDW) captures clinical and operational data from community health centers in a uniform format to facilitate comparisons and analyses over time. There are four categories of data acquired from the clinics: 1) Clinic Operational Data; 2) Clinic Financial Data; 3) Clinic Quality Measures Data; and 4) Patient Level Data. These four data categories are captured and integrated into a common CDW. Downstream data analytics is applied to the data to create customized analytic reports of use to the clinic. Separately, patient analytics provide insights into optimizing care for the CHC’s patient population.

The information flow diagram for the CDW and associated analytics is illustrated in Figure 1.
The conceptualization of the CDW was based on the requirements gathered from all the stakeholders involved by conducting structured interviews with key informants that included clinic management, staff, providers, leadership, patients, as well as IT and QA/QI staff. While the initial choice of the data elements was driven by a narrow purpose to model and simulate clinic’s scheduling system, it soon expanded in scope to serve as a multidimensional repository sourcing information from various domains. The objective is to understand the implications of the clinic operational characteristics on the clinic performance and even patient outcomes over time. For example, the database will allow the study of the impact of scheduling method (such as open access) on appointment compliance (no-shows, cancellations, reschedules), staff utilization, and various patient outcome measures. The system architecture was defined in terms of: data sources, normalized common data model, interoperability requirements, data standards, data acquisition technologies, modes of acquisition, storage in a database server, privacy and security standards set up, data validity checks, data back-up performance and potential to build tools for data management and visualization. Roles and responsibilities were set for each of the above tasks, along with defining the business model.

**Obtaining Clinic and Patient Level EHR Data**

The clinic data category includes: Operational Data (such as, clinic type, certification status, locations, size, patient mix, payer mix, service mix, staff utilization, scheduling system, enrollment system, quality measures and other dimensions of care); and Financial Data (revenue, expense, margins, cash flow, payroll). The clinic operational/financial data was obtained from the audited financial statements of the clinics, IRS Form 990 (annual returns for certain federally tax-exempt organizations) and Uniform Data Systems (UDS) filings. This was supplemented with a structured questionnaire completed through an interview with the clinical manager, followed with workflow observations and opportunistic interviews as appropriate, in the clinic.

The patient level data serves as a backbone of the data warehouse. The data elements selected are those that are digitally captured as part of routine clinical practice and reported to federal agencies. Collecting this data required the design of a common data model, given that we were collecting information from different healthcare facilities using different Patient Management and EHR systems. The common data model helped promote efficient and streamlined collection of the data generated in each of the healthcare centers by organizing data into a standard structure. We developed a common data model that incorporated elements and practices of those used by other large organizations, including: HL7’s C-CDA, PCORnet and FDA’s Sentinel Initiative.
While we explored multiple ways to acquire patient data from the EHR, the one we selected was dependent on the EHR reporting capabilities, IT resources in the clinic, and the data architecture of the CDW. We realized that it was most efficient to obtain the data dictionary/metadata (tables, column names, definition/meaning, data type, predefined values sets/descriptive text) of the EHR where possible and identify the various data elements of our interest. The data elements were selected from well over 3000 tables present in the data dictionaries of the respective CHCs. We then generated SQL queries for each of the individual EHRs and gave them to the clinic staff to execute them on their database server. Once the data was extracted and the de-identification was verified in accordance with HIPAA safe-harbor provisions, the clinic shared the data with us through secure email. The raw data was prepped programatically before populating into our Microsoft SQL Database Server based on the common data model of the CDW and made ready for ad-hoc querying/reporting. Figure 2 shows this process flow of data collection from the clinic EHRs.

The core data element categories included: patient demographics, appointment data, date/time stamps of key events for patient and data flow, insurance information, provider information, medical problem list, encounter diagnoses, immunizations, medications, laboratory data, care plan, procedures, and healthcare referral information.

Results
We now present the results in terms of the common data model developed and a selection of graphs to illustrate the content of the CDW.

Common Data Model
A common data model for the CDW has been built which is guiding the collection of specific data elements. The majority of the data collection exercise, comprising up to five years of EHR and UDS data from all participating clinics, has been completed. For all subsequent data requests (which will be done on a quarterly basis to begin with), the clinic will only need to re-run the queries to append to the current database housed at our end. We mapped the data to the same consistent format (e.g., with the same variable name, precision, and other metadata). By undertaking this step, we created a platform that enables much more rapid analytic capabilities. The data model is expected to grow and evolve over time by including additional data elements that are analytically important.

Note that the data was already de-identified before populating into our database. A pseudo patient ID was used to map the data residing in different tables. In the first round of data collection, many of the raw data tables had to be prepared before loading into our database. Populating free text information was particularly challenging given the existence of delimiters like commas, tabs and new lines. But this was resolved programatically and by specifying the input file format. Future data loading will be much more streamlined. The main tasks in developing the common data model were the selection of appropriate entities/tables and attributes as well as data normalizations based on the three different commercial EHR systems we encountered.
The clinic data warehouse currently stores over 3 million patient-visit records from 2010-14 from the different CHCs and includes detailed information on appointments, prescriptions, lab tests, screening/preventive measures, immunization, and quality measures under Meaningful Use (MU) Stage 2 and UDS. This is a significant volume of rich longitudinal data that will readily lend itself to large-scale data analytics and operational and efficiency studies.

Select Measures on the Data Collected
We present here a partial selection of the key data measures graphically to illustrate the data content.

Figure 4 describes the racial distribution and insurance status of the patient population comprising each of the 7 Indiana CHCs. The distribution clearly shows that the clinics are not homogenous in their patient population and associated needs.

Figure 4: Summary statistics on race and insurance Status

Referring to the plots, it may be noted that the ‘White’ race includes Hispanic/Latinos, since the latter is not defined as a race, but as an ethnicity. In these CHCs, the percentage for Hispanic/Latino ethnicity ranges from a low 1.2% to high 38%. For insurance, a vast majority of them are either uninsured or have Medicaid, with income levels below
200% of the Federal Poverty Line. Yet, there is a sizeable percentage of patients with private insurance in couple of the clinics.

Among our seven partnering CHCs, four use traditional scheduling, while three use modified open access (a combination of same day open slots and traditional/advanced booking). High patient appointment no-show rates and the associated variability in workload due to prevalent practice of overbooking remains a significant challenge in all CHCs. For example, after analysis of the appointment data in the CDW for one of the CHCs, we found that their no-show rate for 2014 was 21% and the appointment cancellation rate was 13%. Providing clinics with easy access to this data and a means to benchmark their performance against other similar organizations can be powerful and create opportunities for design and testing of new technology solutions and positive change.

For another large CHC, we analyzed the visit statistics, in terms of clinic services utilized, and the top 6 disease diagnoses and the associated annual visit frequencies (see Figure 5). It is important to understand the most common medical conditions and service patterns for these predominant conditions in order to focus efforts on study of these problems, and identify potential technology tools that might help improve these challenges. Focus on highly prevalent problems is likely to have the greatest impact on the overall clinic efficiency and other operational and financial measures.

![Figure 5: Statistics on services utilized and medical conditions with associated visits at a large CHC in Indiana](image)

**Discussion**

We have established a novel information exchange/clinic data warehouse that will enable data-driven research, technology development and testing, and informed decision making over the years. We have presented only a small selection of the data in the repository for illustration and plan a number of studies to better investigate its use and applications. We believe this approach can provide an efficient and effective means for CHCs to leverage data to improve patient access and the quality of care delivered to underserved patients.

**Academic Research and Benefits to CHCs**

There are innumerable research opportunities that can come from analytic study of the CDW. These may be broadly classified as being, descriptive, predictive, and prescriptive analytics. A few examples under, ‘descriptive analytics’ are: development of enriched careflow/process models based on EHR time-stamps; ranking clinics based on adherence to evidence based best practices; and peer comparisons at the state and national level. Under ‘predictive analytics’ the plan is to build machine-learning models of patient no-show behavior and treatment compliance; to perform agent-based modeling and simulation of clinic operations to improve efficiency; and to perform longitudinal network analysis and risk stratification of diseases. Note that although the patient data in the CDW is deidentified, a pseudo patient ID will still allow us to conduct patient-centered longitudinal studies across

---

1 Enabling services in the pie chart include: case management, outreach, patient/community education, transportation, eligibility assistance, and interpretation services.
2 Mental Disorders in the graph is cumulative of 4 mental disorder types: Depression, Anxiety and PTSD, Attention Deficit Disorder, and Other Mental Disorders.
care settings. Assessing the impact of the affordable care act (ACA) on access to care will be of particular and timely relevance.

Under ‘prescriptive analytics’ we intend to incorporate findings from predictive analytics into standalone software applications or embed within patient management systems to provide personalized decision support based on patient profile and clinic characteristics. The data can be further enriched by including external variables, such as data from community/geo-spatial information systems. We also are particularly excited about the opportunity for research and development of new technologies for patient engagement and management in the community, as well as integration of disparate care providers and care sites (inpatient, outpatient, community, “minute” clinics, prisons, other) as part of a care management team. The community clinics stand to benefit from such analytics since they lack the infrastructure that is usually available to academic medical centers.

Ongoing Work
We are refining the design of a data-driven clinic dashboard that will be linked to the CDW. Figure 6 shows the mock-up of this dashboard under development, which initially will focus on measures of access to care and efficiency. It has a selection of key metrics across different categories (operational, quality, and financial) and has been derived from key informant interviews. It will allow clinic management to visually assess the impact of their operational changes (such as scheduling, staffing level, and policy changes) on clinical and non-clinical outcomes on a periodic basis.

<table>
<thead>
<tr>
<th>OPERATIONAL-1</th>
<th>OPERATIONAL-2</th>
<th>QUALITY</th>
<th>FINANCIAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-Show Rates</td>
<td>New Visits</td>
<td>QA/QI 50th/90th Percentile</td>
<td>Billed Visits</td>
</tr>
<tr>
<td>Checked Out Visits</td>
<td>Phone Abandonment</td>
<td>Continuity of Care</td>
<td>Charge Lag</td>
</tr>
<tr>
<td>Employee Retention</td>
<td>Booking to Appt. Time</td>
<td>Patient Satisfaction Scores</td>
<td>Enabling Services Coded</td>
</tr>
<tr>
<td>Provider Productivity</td>
<td>Door to Door Time</td>
<td>Population Health Management</td>
<td>Patient Payments</td>
</tr>
</tbody>
</table>

Figure 6: Mock-up of a community clinic dashboard with associated metrics of relevance

We are also working on a secure web portal to serve as an interface for clinic data submission and running queries by researchers. The CDW portal will allow users to explore, interact with, and export data through charts, maps, reports, analyzer and locator tools, data downloads and data services, and widgets. Finally, we are developing strategies for interacting with technology developers and entrepreneurs, in order to test the feasibility of new tools and prototypes.

Challenges and Strategies
There were various challenges encountered in acquiring the data from the clinics. For most clinics, this was the first time they received such a request to extract EHR data. Thus far, the clinics primarily ran reports to fulfill government requirements, in tandem with manual charting. Even though the clinics lacked adequate time or
resources for learning and drawing insights from their data, they all readily recognized its importance to support their evidence-based quality improvement efforts, and be in compliance with the upcoming MU Stage 3 requirements and intent to seek PCMH certifications.

In terms of strategies, we sought to engage meaningfully with all the clinics and presented our value proposition to the clinic leadership. We interviewed the clinic staff and patients to understand their challenges and learn from their innovations, coupled with onsite workflow observations of key processes. We were able to obtain the EHR data dictionary from 2 different EHR systems (covering 5 out of the 7 clinics). These EHRs were based on client-server architecture. The data dictionary allowed us to identify all the data elements of interest to us and create the SQL queries to be executed on their database servers. This made it less onerous on the clinics as they were able to run the provided queries and share the extracted data with us in a secure mode. The remaining 2 CHCs were using a cloud-based EHR with no direct access to the database server. We had to work rather closely with them to pull the data from their web-based reporting tool.

Study Limitations
There were a few limitations to our study. First being, that since the data dictionaries were from three different commercial EHR vendors, the common data model may not reflect the universe of EHRs in the ambulatory care space. Furthermore, the data warehouse architecture has been partially influenced by the unique characteristics of CHCs in the state of Indiana and therefore may not reflect the particulars of other CHCs in the US. The model may be even less representative of other primary care practices who are not dedicated to meeting needs of the underserved. Finally, the design of the architecture and resource constraints imposes certain practical limitations on the frequency of data collected from the clinics, which in turn would define the type of analytics and decision support that can be performed.

Conclusions
We have described a multidimensional clinic data warehouse for CHCs spread across Indiana and one that integrates operational, financial, quality and patient data. We chose a centralized data architecture that is relatively easy to implement and sustain, yet is rich in scope and content, and relevant to the participating clinics. The goal is to enable free standing CHCs to participate in this data warehouse, in the same way that large hospitals participate in HIEs and benefit from large scale healthcare data analytics to improve overall clinic operational performance and achieve better scores on the ever evolving clinical quality measures. Developing strategies for interacting with technology developers and entrepreneurs, in order to test the feasibility of new tools and prototypes should speed innovation and dissemination of effective IT. Ultimately, we believe that providing the CDW will help CHCs better meet the needs of their patients.

Acknowledgement
The author would like to thank and acknowledge PCORI, which has funded this 3 year, $2M grant award [IH-12-11-5488; ‘Improving Access to Care and Efficiency of Healthcare Systems for Underserved Patients’, PI: Doebbeling]. The author would also like to acknowledge the staff at the participating clinics for their support and the IU School of Informatics and Computing for providing the necessary technology infrastructure for the implementation of this clinic data warehouse.

References

Creating Shareable Clinical Decision Support Rules for a Pharmacogenomics Clinical Guideline Using Structured Knowledge Representation

Margaret K. Linan, MPH\textsuperscript{1}, Davide Sottara, PhD\textsuperscript{1,2}, Robert R. Freimuth, PhD\textsuperscript{2,3}

\textsuperscript{1}Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ, \textsuperscript{2}Office of Information and Knowledge Management and \textsuperscript{3}Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract
Pharmacogenomics (PGx) guidelines contain drug-gene relationships, therapeutic and clinical recommendations from which clinical decision support (CDS) rules can be extracted, rendered and then delivered through clinical decision support systems (CDSS) to provide clinicians with just-in-time information at the point of care. Several tools exist that can be used to generate CDS rules that are based on computer interpretable guidelines (CIG), but none have been previously applied to the PGx domain. We utilized the Unified Modeling Language (UML), the Health Level 7 virtual medical record (HL7 vMR) model, and standard terminologies to represent the semantics and decision logic derived from a PGx guideline, which were then mapped to the Health eDecisions (HeD) schema. The modeling and extraction processes developed here demonstrate how structured knowledge representations can be used to support the creation of shareable CDS rules from PGx guidelines.

Introduction
In this work, we propose a consistent and reproducible methodology to generate platform-independent, shareable representations of PGx guidelines using existing and emerging standards. We modeled a published PGx guideline using UML activity diagrams, the HL7 vMR data model, and standard terminologies. The aim is to use these standards to build a structured, human-readable document and, ultimately, a shareable, machine-readable artifact. The motivation for developing a standards-based knowledge representation process that yields sharable policies is twofold. First, no commonly accepted, standard methodology currently exists for the extraction of CDS rules from PGx guidelines. Second, the creation of shareable CDS rules from PGx guidelines will provide clinicians and implementers with standards-based decision logic that can be used as a starting point for the design of evidence-based care delivery interventions.

Adverse drug reactions (ADR) cause more deaths than AIDS and diabetes, making it the fourth major cause of death for over a decade\textsuperscript{1}. Approximately 100,000 deaths and more than 2 million acute ADRs occur annually\textsuperscript{1}. Some of these ADRs are caused by variations in genes associated with drug metabolism or mechanism of action. Many of these gene-drug interactions are clinically actionable. Groups such as the Clinical Pharmacogenetics Implementation Consortium (CPIC) develop clinical guidelines that provide clinicians with information about known gene-drug interactions. These guidelines also provide clinicians with recommendations that help guide drug therapy\textsuperscript{2,3,4}.

Like most clinical guidelines, pharmacogenomics (PGx) guidelines are written for clinicians and are not available in computable formats. While CPIC is beginning to include some structured text within their published guidelines, formal representations of the knowledge are not available, which hinders the extraction and implementation of CDS rules\textsuperscript{1}. In addition, there are several other factors that complicate the implementation of PGx CDS. One of the main issues is the burgeoning amount of biomedical knowledge; there are currently 166 known gene drug interactions on the FDA website and 21 PGx guidelines have been published, and these numbers are growing rapidly\textsuperscript{5,6,7,8}. Another factor is the very costly and time consuming process that is required to extract PGx knowledge from narrative guidelines, implement them as CDS rules in electronic medical record (EMR) systems, and then maintain those rules as biomedical knowledge evolves. Also, there are differences in how a clinical practice guideline can be implemented as a CDS rule, due to differences in the interpretation of the guideline and local workflows. It may be possible to reduce the cost of rule development and improve the consistency of CDS implementations by sharing CIGs, but currently there are no best practices or knowledge bases available to support the authoring of CIGs\textsuperscript{9}. Therefore, a scalable approach is needed to extract, render and implement CDS rules from PGx guidelines. To accomplish this, common semantics and a platform-independent (vendor-agnostic) syntax is required.
Standardized data models and terminologies reduce variability in the representation and interpretation of clinical data by providing common semantic meanings for concepts and terms. For example, the HL7 virtual medical record (vMR) data model was developed to represent clinical data from an EMR in a platform-neutral manner, which can be used for CDS. Similarly, standard terminologies, such as SNOMED CT, RxNorm and LOINC, provide common definitions for clinical terms, drugs, and lab tests, respectively. While the use of common data models and standard terminologies increases comprehension and reduces semantic ambiguity in shared data, they are not intended to represent the workflow and decision logic inherent in CDS rules. To represent these aspects, different standards are needed.

Several modeling frameworks, such as the Guideline Elements Model (GEM) and the Guideline Interchange Format (GLIF), have been developed to represent the content of clinical guidelines. However, no existing framework has achieved broad adoption and the variety of approaches, formats and limitations of existing modeling tools prevents artifacts generated using these frameworks from being shareable in practice. Despite decades of work in this area, there remains a dearth of applications that can support the authoring of platform-independent, shareable CIGs for use in CDS systems. To address this challenge, in 2012 the Office of the National Coordinator (ONC) for Health Information Technology supported the creation of a harmonized format, known as the Health eDecisions (HeD) interchange format. The HeD standard defines a common metamodel for metadata, actions, events, and conditions as well as an expression language. HeD also recommends an XML-based serialization format and provides a schema for the validation of compliant documents. Together, the HeD model and its schema were designed to represent the workflow and decision logic of interventions such as CDS rules, order sets and documentation templates.

To complement the HeD schema, in the context of the SHARPc-2B project, a multi-institution team developed a more formal representation of the HeD metamodel, expressed in the form of an OWL ontology, and a standards- and model-driven application for the authoring CDS clinical knowledge artifacts. The HeD editor has several features that can potentially make it an ideal part of the process for modeling and testing the clinical decision logic for PGx guidelines. Specifically, it is compatible with several existing and pre-existing standards, has an intuitive user interface that can be used by knowledge engineers and non-technical staff, and uses semantic web technologies.

In this work, we modeled a published PGx guideline using UML activity diagrams, the HL7 vMR data model, and standard terminologies. We then utilized the HeD editor to render the modeled guideline in HeD syntax. This work builds on our previous efforts to create platform-independent, standards-based, shareable representations of PGx guidelines and is another step towards the development of a generalizable approach that can be broadly applied to other PGx guidelines.

**Methods**

Published CPIC guidelines were reviewed and evaluated in terms of both the complexity of logic and the availability of reference implementations of corresponding CDS rules. For this project, we sought a PGx guideline that contained straightforward decision logic that was based on unambiguous PGx data (genotype and phenotype). We also prioritized guidelines that had been implemented as local, non-shareable CDS rules, which would provide additional clinical context over that which was included in the published guideline.

A stepwise process was developed to transform a human-readable PGx guideline into a computable HeD artifact (Figure 1). The process began by carefully analyzing the selected PGx guideline and rigorously defining all of the concepts, both explicit and implicit, that were relevant to the implementation of a PGx CDS rule. This included concepts that pertain to the clinical context of the guideline, the relevant EMR data, and terminology. The process was informed by knowledge of actual CDS implementations of the selected guideline.

Figure 1. The workflow for modeling a PGx clinical guideline and expressing the information in HeD syntax.
Once relevant concepts were identified within the PGx guideline, candidate reference standards were reviewed and evaluated to determine which might best represent those concepts. Specifically, we evaluated the HL7 vMR, HL7 RIM, and LS DAM, to represent the clinical data for this PGx guideline\textsuperscript{21,22}. We also evaluated the RxNorm, NDF-RT, SNOMED CT, and LOINC standard terminologies. The standards that were selected were used in the next step, when the PGx guideline itself was formally modeled and rendered in HeD syntax.

The parts of the PGx guideline that were targeted for modeling were the decision tree (which contained the recommended clinical workflow), the therapeutic recommendations, and the table for the genotype-informed decision logic, as the content in these sections represented the information that would comprise CDS rules. The information from these sections was augmented with knowledge of the corresponding CDS rules at Mayo Clinic, which provided additional clinical context about real-world PGx interventions. Collectively, these data were used to create a high-level model, which was expressed using UML class and activity diagrams to describe the logic contained within the PGx guideline. To reduce its complexity, the high-level model was split into logical modules that each contained a small number of decision points.

Each of the modules were analyzed in detail to identify the data elements, terminology concepts, and functions (e.g., data retrieval from the EMR, result processing) that would be necessary for implementation as a CDS rule. This information was then used to determine which entities (e.g., classes, attributes, and terms) from the selected reference standards would be needed to represent the entities within each module. Elements that represented the entities from the reference standards were added to the UML model and associated to their respective classes from the PGx guideline. Lastly, activity diagrams were generated to refine the computational steps within each workflow. Together, the UML models captured the entities, workflow, and behavior contained within the PGx guideline, but an additional method was needed to represent the logic of the CDS rules.

The logic components of the PGx guideline were expressed in pseudocode, which described the flow of data between each step and represented the functions that would be needed to implement each CDS rule. The pseudocode was based on an early version of CQL, and was used to define variables and create if-then conditional logic blocks that evaluated clinical data represented by vMR classes and attributes.

Finally, the model and pseudocode were used to render the guideline in HeD syntax, referencing the data and terminology standards (Figure 2). The HeD editor was used to ensure the resulting artifacts complied with the HeD schema.

![HeD Pseudocode (Text)](Figure 2. The HeD editor was used to generate structured rules from an unstructured PGx guideline, which was modeled in UML. Data and terminology standards provided common semantic meaning, and the logic of the workflow was expressed using the HeD schema.)
Results

The CPIC guideline for HLA-B genotype and abacavir dosing was chosen because of its straightforward decision logic. In particular, the HLA-B genotype is expressed as a boolean result (HLA-B*57:01 allele present/absent) that determines whether or not abacavir is contraindicated. The core recommendations from the guideline are summarized in Table 1. This guideline has been implemented as a CDS rule at Mayo Clinic.

Table 1. Core recommendations from the HLA-B/abacavir guideline.

<table>
<thead>
<tr>
<th>HLA-B*57:01 Genotype</th>
<th>Clinical Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-B*57:01 allele absent</td>
<td>Abacavir is not contraindicated</td>
</tr>
<tr>
<td>HLA-B*57:01 allele present</td>
<td>Abacavir is contraindicated</td>
</tr>
<tr>
<td>HLA-B*57:01 status unknown</td>
<td>Order genotype test</td>
</tr>
</tbody>
</table>

The HL7 vMR model, which was developed in part to support CDS use cases, was selected to represent clinical data elements. The vMR includes classes that represent the concepts of patient, drug order, ADR, test order and result, and communication events (e.g., CDS alerts). The RxNorm, SNOMED CT, and LOINC standard terminologies were chosen to provide coded concepts for drugs (e.g., abacavir), ADRs (e.g., adverse reaction to abacavir), and lab tests (e.g., abacavir genotype), respectively. While some of the coded concepts required for this project existed within the selected terminologies, several concepts were missing. Specifically, SNOMED CT did not contain a pre-coordinated term for an abacavir-related ADR but it would be possible to express this concept through post-coordination using "Antiviral drug adverse reaction" (292826004), "Causative agent" (246075003), and "Abacavir" (387005008). In addition, LOINC did not contain codes for the HLA-B*57:01 genotype test or results. Placeholder concept codes were used for the genotype test and results when rendering the rule in HeD syntax.

The HLA-B/abacavir guideline was modeled as five inter-related logical modules that represent the clinical workflow, therapeutic recommendations, and decision logic that would be necessary for implementation of the guideline as a CDS rule (Rules A-E, see Figure 3). These modules contain models that describe the entities, attributes, concepts, workflow relationships, and processing functions to search the patient’s record for a documented ADR (Rule A) and genetic test report (Rule B). CDS interventions are modeled in Rules C, D, and E.

Figure 3. Overview of the five inter-related rule modules that were generated for the HLA-B/abacavir guideline

When the CDS system detects an order for abacavir Rule A is triggered, which queries the EMR to see whether the patient has a documented adverse reaction to abacavir. If an adverse reaction is found then a CDS alert fires, instructing the clinician to cancel the drug order. If no adverse reaction exists then Rule B is triggered, which queries the patient’s record for HLA-B*57:01 genetic test results. If no results are found then Rule C is triggered, which recommends the clinician order the genetic test (if it has not already been ordered) or delay the drug order.
until the test results are available (if the test has been ordered and results are pending). If test results are found, Rule D is triggered, which examines the report for coded genotype results and either fires an alert to the clinician (HLA-B*57:01 allele present, abacavir is contraindicated) or allows the order to proceed (HLA-B*57:01 allele not present). If Rule D is triggered but coded genotype results are not available then Rule E is triggered, which examines the report for coded phenotype results (e.g., high/low risk of abacavir hypersensitivity). If coded phenotype results are available, the system acts accordingly (see Rule D); otherwise, the test results have not been codified and the CDS system will advise the clinician to manually review the genetic test results report.

The entities in each rule were mapped to the corresponding entities from selected data and terminology standards. This step resulted in the identification of the vMR classes and attributes that would be needed for each rule. For example, the classes identified for Rule A included EvaluatedPerson, SubstanceAdministrationOrder, AdverseEvent, and CommunicationProposal (Figure 4).

Some of the attributes have a datatype of CD (concept descriptor), which reference coded concepts from terminologies that can be defined by the implementer. The standard terminologies selected for this project were used, when possible, as sources for terms that were considered to be critical for the rendering of the rule in HeD syntax. For example, “abacavir” can be represented by RxNorm RxCUI 190521, which can be used in Rule A for both SubstanceAdministrationOrder.substanceGenericCode and AdverseEvent.adverseEventAgent (Figure 4).

**Figure 4.** Clinical workflow for Rule A, with concepts identified and mapped to HL7 vMR classes. Not all attributes are shown.
Each of the five workflows were refined into UML activity diagrams, which more carefully defined the actions required at each step of the workflow (Figure 5). The activity diagrams represented the final stage of UML modeling and, with the workflows described above, informed the generation of pseudocode.

Figure 5. An activity diagram for Rule A, showing the activities that need to be performed by the CDS system. These process steps were later expressed in HeD syntax.

The penultimate step in this process involved creating pseudocode to outline the underlying processes involved in the retrieval and analysis of clinically relevant information from the EMR (see Figure 6 for an example). The pseudocode facilitated the transformation of the information specified in UML into a more pragmatically defined structured language while still retaining some aspects of human readability.

```plaintext
valueset "Abacavir" = Valueset( x.y.z )
let AdverseReactionsToAbacavir =
    [ Occurrence of AdverseEvent : adverseEventAgent in "Abacavir" ]
let hasAdverseReaction = exists AdverseReactionsToAbacavir
let Message = 'Consider cancelling the order for patient ' + name.family + '...'
```

Figure 6. An example of pseudocode corresponding to a portion of Rule A. References to the vMR are shown in red, while defined variables and expressions are shown in blue.

An initial HeD rendering of Rule A was created using the HeD editor and the knowledge artifacts previously generated (e.g., UML models, pseudocode). This rendering included references to classes and attributes from the vMR, concept codes (or placeholders) from standard terminologies, and the workflow logic expressed using the HeD schema (Figure 7).
Figure 7. An example of the HeD syntax corresponding to a portion of Rule A. References to the vMR are shown in red; defined variables, expressions, and values are shown in blue; elements from the HeD model are shown in green.

Discussion

In this project we sought to develop a standards-based methodology to generate platform-independent, shareable representations of PGx rules. We modeled the published HLA-B/abacavir PGx guideline using UML class and activity diagrams, the HL7 vMR data model, and standard terminologies. These models were used to create pseudocode representations of the rules, which were ultimately rendered in HeD syntax. Thus, we began with an unstructured, human-readable clinical guideline and produced a shareable, machine-readable artifact.

This is the first time that PGx clinical guidelines have been formally modeled in UML with the specific aim of rendering the extracted rules in a platform-independent language, such as HeD. The use of UML to model the rules contained in the guideline facilitated the identification of data elements and coded concepts, and the subsequent mapping of those entities to standards. The UML modeling process also helped to provide a clear representation of key decision points in the rule and clinical workflow, and made explicit the computational tasks that need to be performed during the execution of the rules, such as data retrieval and processing. We anticipate this modeling approach will scalable and generalizable to other PGx guidelines. A review of other CPIC guidelines revealed variations in genetic test interpretations (e.g., phenotype) and the complexity of therapeutic recommendations, so it will be important to demonstrate this method can be used for a variety of PGx guidelines.
The HL7 vMR model, which was developed in part to support CDS use cases, was selected to represent clinical data elements. The vMR model was chosen, in part, because it provided the best coverage overall of the clinical concepts that were expressed in the PGx guideline and rules. We did not require the full expressive power and complexity of the HL7 RIM, of which vMR is a derivative, and other competing standards like FHIR were not sufficiently mature at the time this work was performed. We also did not consider logical or more detailed clinical models, such as the ones delivered as Clinical Element Models (CEMs) or OpenEHR archetypes as we decided that while we would likely find models for the more traditional concepts used in the rule, those sources did not provide PGx-specific models. For example, genetic test results were represented in the vMR as Observations, which is used for generic lab tests, but a refined domain model may be required to more fully capture the nuanced semantics of genomics data, including the concepts of allele, haplotype, copy number, and predicted phenotype. The HL7 Clinical Genomics Working Group is currently working in this area and we will evaluate the outcome of those projects when the work is complete.

The standard terminologies that were chosen for this project had variable coverage of the concepts contained in the PGx guideline. Both RxNorm and SNOMED CT contained a concept for "abacavir"; we chose to use the term from RxNorm as it is more likely to be supported by pharmacy systems. SNOMED CT did not contain a single term to represent a documented ADR for abacavir, but as noted above it would be possible to express this through post-coordination. Since the vMR model provides an attribute to specify the cause of an adverse event, though, this attribute was used with the code for "abacavir" rather than the more complex approach of using post-coordinated terms, since the class and the attribute already convey part of the semantics.

We were not able to locate an entry in LOINC that represented the HLA-B*57:01 genetic test or the results from that test; this was not a surprise, as standardized terminology is a known gap in the PGx domain. CPIC is leading a terminology harmonization effort that will provide more consistency among its guidelines and it is likely that the terms and value sets resulting from that project will be proposed for inclusion in LOINC. As those terms were not available at the time of this study, we used placeholder codes for these concepts in the HeD rendering of the rules.

The data and terminology standards provide common semantics that enable the sharing of rules derived from PGx guidelines. The actual implementation of CDS rules would require a site-specific mapping from the standard-based representation to local data models and terminologies, which reflects differences in where in the EMR data is stored and how it is coded. Similarly, while HeD is designed to represent the workflow and decision logic of interventions in a platform-independent manner, currently there is no consolidated way to execute HeD syntax or automatically transform it to an EMR-specific language beyond initial pilot implementations. These limitations represent significant opportunities for further research and tooling development.

Conclusion

In this work we demonstrated a method for modeling the HLA-B/abacavir PGx guideline in UML, which can represent the clinical workflow and decision logic as standards-based structured knowledge, and we illustrated how these models can be used to aid rendering the rules in the HeD schema. We anticipate this process will be generalizable to other PGx guidelines, although several important gaps in reference standards and tooling need to be addressed. This approach will enable the creation of sharable, platform-independent knowledge artifacts that may facilitate the implementation of and consistency among PGx CDS rules.

Acknowledgments

The authors would like to thank Pooja Raghani for reviewing the modeling protocol. This work was funded in part by the NIH/NIGMS (U19 GM61388; the Pharmacogenomics Research Network) (RRF).
References


Analyses of Merging Clinical and Viral Genetic Data for Influenza Surveillance

Daniel Magee¹, Rachel Beard¹, Matthew Scotch PhD, MPH¹
¹Arizona State University, Tempe, AZ, USA

Abstract
The annual influenza vaccine is one of the most common public health interventions and is universally recommended for all individuals older than six months. Vaccine composition depends on viruses circulating over the past flu season and are estimated to be the most prevalent and representative strains in the current season. Here, we use clinical data outfitted with viral genetics to characterize confirmed influenza cases from the past two flu seasons and genetically compare them to the strains that they were vaccinated against that year. We show that case similarities to vaccine strains differ by geographic region and that the vaccines appear to have different levels of effectiveness by region. This study demonstrates the value of merging viral genetics with clinical data. Further research is needed to formally evaluate whether this improves biosurveillance efforts and enhances efficacy of influenza vaccines.

Introduction
One of the cornerstones of public health improvements over the past several decades is widespread vaccination against potentially threatening diseases such as tetanus, measles, mumps, rubella, and hepatitis. Of particular note, influenza has caused a number of pandemics as a result of novel strains circulating in the general population, such as the 1918 H1N1 Spanish influenza which caused an estimated 20-50 million deaths worldwide¹. More recently, global influenza infection rates are estimated to include 5-10% of all adults and 20-30% of children annually during a non-epidemic year² amounting to 250,000-500,000 deaths according to a 2014 a World Health Organization (WHO) report³. In an effort to curb these rates, programs to vaccinate a large portion of the population are carried out each year in countries such as the United States, resulting in annual vaccination rates falling between 39.4-42.2% of the population from 2009-2014⁴. However, efficacy of these vaccinations can vary significantly year to year by region, demographics, and time of vaccination⁵,⁷.

Influenza vaccines to counter detrimental effects on public health were introduced in the 1940s, initially including at least one influenza A and B strain⁸. Each vaccine uses a dead or weakened strain of virus to promote the body’s immune response and help protect against an actual infection with little to no risk to the patient⁹. The Centers for Disease Control and Prevention (CDC) provides a recommended vaccination schedule for both children⁹ and adults¹⁰. Included in these schedules is the annual influenza vaccine which comes in two forms: inactivated influenza vaccine (IAV) and live, attenuated influenza vaccine (LAIV)¹¹ and are administered depending on the age and status of the patient. These vaccines are generally trivalent and can protect against three of the most common types of influenza: A/H1N1, A/H3N2, and B. It is recommended that all individuals at least six months old receive a vaccination annually¹² in time for the beginning of the flu season, which falls between weeks 40-20 in the Northern Hemisphere. In preparation for this demand, 150-160 million vaccines are annually produced for use in the United States.

The vaccines are created based off of recommendations by the WHO and are a result of a continuous process analyzing data and viruses that circulated during the previous flu season. There are 130 national influenza centers (NIC) in 101 countries¹³, including four in the United States. These centers are tasked with analyzing clinically-confirmed influenza cases by a variety of genetic and antigenic tests and determining the strains that are the most representative or novel. These strains are then forwarded to one of five WHO Collaborating Centers for consultation and further testing¹³. The additional testing involves inoculation into ferrets for antisera analysis, human serology studies, analysis of antiviral resistance properties, and phylogenetic clade analysis. After taking these results into consideration, primary selection of the H1N1, H3N2, and B viruses are made and initial vaccines are created for further analysis; the H1N1 and H3N2 strains are created by classical reassortment while the wild type B virus is used because there is no growth advantage for its reassortment. These vaccines are used to gather ferret antisera once more to ensure that growth properties and antigenic content are sufficient and that the antigens created are identical to those of the wild type viruses. In the event that the antigenic state or growth properties are insufficient, the WHO may decide to try to utilize all wild type viruses

1995
or may select entirely new strains. Once sufficient vaccines have been created and demonstrated, the WHO will make its recommendations for the upcoming flu season and these recommendations apply to the entire Northern or Southern Hemisphere; however each country generally has a local agency approve the recommendations or choose different, but antigenically similar, strains. In the United States, final approval is provided by the Food and Drug Administration (FDA). Mass production of vaccines begins after final virus selection is made, usually February through May, in order to have the massive amount of required stockpiles ready for the beginning of the vaccination period in October.

The most important takeaway from the selection and production process is that the strains selected for inclusion in the vaccine reflect highly educated guesses of the strains that will be the most representative in the upcoming flu season. Vaccine effectiveness depends on the actual strains circulating when the flu season arrives, the number of individuals that receive the vaccination, and several other factors. Early estimates of the 2014-2015 flu season in the United States indicate overall vaccine effectiveness at just 23% due to genetically drifted H3N2 strains in circulation. The same vaccine is generally distributed throughout the entire United States, with the exception of some quadrivalent vaccines that protect against an additional influenza B virus, and do not take into account locally circulating viruses in real time during the flu season. This enables phenomena like the current flu season where antigenically divergent strains predominate and reduce efficacy of the vaccine. Global geographic dispersal has been implicated as the main mechanism by which the influenza A virus survives and evolves into antigenically divergent strains between seasons. This demonstrates that geography can be a causal factor toward the prevalent strains during a flu season and led us to explore the use of clinical data from confirmed influenza cases to analyze genetic similarities between the cases and the viruses that were vaccinated against. These types of data may be used to help clinicians decide on treatment options vaccines that may be available to administer to patients in their area, and help governing bodies of public health better determine strains to target for vaccination at a more granular level. For instance, this type of analysis would be useful for physicians at the local level to prescribe drugs potentially more effective at treating patients based on the strain of influenza they are infected with. Additionally, given the availability of a complete data sets encompassing both genetic and clinical characteristics of a flu isolate and its associated host, this would allow tracking of effectiveness of vaccines for particular regions in conjunction with trends in strain resistance factors.

The CDC oversees a federal influenza surveillance program that includes participation from state health agencies. The current initiative includes subtyping, antigenic characterization, and antiviral resistance testing. There is also information on influenza-associated hospitalizations and outpatient influenza-like illness. Beyond Federal reporting, some individual healthcare facilities have begun to share both clinical and viral genetic data with the researcher community. For example, the influenza research database (IRD) contains a relatively new feature for facilities to share clinical data from influenza-related hospitalization as well as the sequences of the virus itself. Currently sixteen distinct facilities are participating including Beth Israel Deaconess Medical Center, Children’s Hospital Boston, the University of Rochester Medical Center, and the National Research Center in Egypt.

In this study, we analyze overall and geographic distribution of genetic similarities between the cases and vaccines and demonstrate the potential of combining clinical and viral genetic data to explore trends of influenza epidemics.

Methods

We collected clinical data for influenza cases from the IRD. We used the Human Isolates with Clinical Metadata function and searched for all data from Week 40-20 for each of the 2012-2013 and 2013-2014 flu seasons in the United States. Of the returned results, all categories of data were selected and downloaded. These included: GenBank database accession ID, age of the patient, gender, vaccination status and date, presence of fever and other clinical symptoms, as well as genetic data about the virus such as markers for antiviral resistance (adamantine or oseltamivir). We discarded any record that was not influenza A subtype H1N1, H3N2, or type B. Of the remaining cases, we discarded any record that did not contain a specific strain or sequence accession number. We used the accession IDs in our result set to search GenBank and download the hemagglutinin (HA) sequence for each case. We discarded all records containing a partial sequence and this resulted in the final dataset. In Figure 1, we show a flowchart of the data collection and preprocessing.
Figure 1. Flowcharts resulting in the cases we used for analysis in this study. A) 2012-2013 flu season with data representing 81.2% of all cases. B) 2013-2014 flu season with data representing 79.9% of all cases.

In order to compare the sequences with those contained within the vaccine, we collected a strain recommended by the WHO for inclusion in the respective flu seasons in the Northern Hemisphere from the IRD. The WHO recommends “like-virus” strains (e.g. A/Brisbane/59/2007(H1N1)-like virus) rather than a wild type strain, rather than specific strains, and the strains vaccinated against may differ between vaccine manufacturers. This makes it impossible to conclude exactly which strains each individual is vaccinated against in a set of deidentified data. For this reason, we collected a random HA sequence from GenBank matching the suggested strains within the IRD. In Table 1, we describe these strains.

<table>
<thead>
<tr>
<th>WHO Recommendation</th>
<th>2012-2013 Flu Season</th>
<th>2013-2014 Flu Season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain</td>
<td>H1N1</td>
<td>H3N2</td>
</tr>
<tr>
<td>Accession ID</td>
<td>CY121680</td>
<td>KJ942680</td>
</tr>
<tr>
<td>A/California/07/2009(H1N1)</td>
<td>A/Victoria/361/2011(H3N2)</td>
<td>B/Wisconsin/01/2010</td>
</tr>
<tr>
<td>CY121680</td>
<td>KJ942680</td>
<td>CY115183</td>
</tr>
<tr>
<td>A/California/07/2009(H1N1)</td>
<td>A/Victoria/361/2011(H3N2)</td>
<td>B/Massachusetts/02/2012</td>
</tr>
<tr>
<td>KC891816</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. WHO Recommendations for Composition of Influenza Vaccines in the Northern Hemisphere by flu season and their corresponding GenBank Accession ID. These Accession IDs correspond to the specific HA sequence we used for analysis in this study.

We aligned sequences for each virus subtype from each season with the sequences representing the subtype contained in that year’s vaccine using Geneious Pro v.6.1.6 (Biomatters Ltd., Auckland, New Zealand) and MAFFT v7.017. We utilized the default parameters (Algorithm = auto, Scoring Matrix = 200PAM / k = 2, Gap Open Penalty = 1.53, Offset Value = 0.123). Post-alignment, we measured genetic distance in the number of single nucleotide polymorphisms.
(SNPs) from each sequence to the vaccine and translated the aligned sequences with Geneious to determine the number of nonsynonymous SNPs. We appended these distances to our clinical data file. In order to quantify the impact of geography among the cases, we appended a column in the clinical data file with the United States Department of Health and Human Services (HHS) Region corresponding to the state from which the case was isolated. There are ten HHS Regions\(^{25}\) in total representing each of the 50 United States and several territories including American Samoa, Guam, and others.

**Results**

The results we obtained focus on genetic variation among confirmed flu cases and the vaccinated strains as well as a metric of vaccine effectiveness among cases in both the regional and overall scopes. In Figure 2, we show the number of SNPs compared to the strain in which it was vaccinated against.

![Number of SNPs in Confirmed Cases Compared to Strain in Vaccine](image)

**Figure 2.** The number of post-alignment SNPs in our cleaned set of cases compared to the strain in which it was vaccinated against (Table 1). The blue and orange bars represent the average synonymous and nonsynonymous SNPs per virus per year, respectively, with error bars as standard deviations for n cases, listed above.

After eliminating cases with unknown vaccination status, we determined that 49.2% and 50.2% of our cases were vaccinated for the 2012-2013 and 2013-2014 flu seasons, respectively. The lack of control individuals here makes calculation of vaccine effectiveness and efficacy\(^{26}\) impossible. We summarize these data in Table 3.

<table>
<thead>
<tr>
<th>Vaccination Status</th>
<th>2012-2013 Flu Season</th>
<th>2013-2014 Flu Season</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H1N1</td>
<td>H3N2</td>
</tr>
<tr>
<td>Yes</td>
<td>44.0%</td>
<td>24.6%</td>
</tr>
<tr>
<td>No</td>
<td>28.0%</td>
<td>22.7%</td>
</tr>
<tr>
<td>Unknown</td>
<td>28.0%</td>
<td>52.7%</td>
</tr>
<tr>
<td>C(</td>
<td>VS+,Y)</td>
<td>61.1%</td>
</tr>
</tbody>
</table>

**Table 2.** Overall vaccination rates among analyzed cases. Unknown includes cases with unknown or missing statuses. C\(|VS+,Y) can be read as “cases with a vaccination status of yes given a known vaccination status”.

Also of interest was the variation of vaccination status among the HHS Regions in the United States in order to capture
geographic trends and compare them with genetic features. We summarize the vaccination statuses of our cases by region and year in Table 3. We visualize these data in Figures 3 and 4 by showing the average number of SNPs per case, separated by strain, for each region. Background color of the regions indicates the $C|VS+,Y$ column shown in Table 3.

<table>
<thead>
<tr>
<th>Region</th>
<th>2012-2013 Flu Season</th>
<th>2013-2014 Flu Season</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>127</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>141</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>89</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>43</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>19</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3. Summary statistics of vaccination status of each patient by HHS Region. UN refers to individuals with a status of “unknown” or “N/A”. $C|VS+,Y$ is the proportion of cases where the individual was vaccinated to the number of cases with a known vaccination status, expressed as a percent (previously defined in Table 2).

Figure 3. 2012-2013 flu season by HHS Region, indicated in boldface. Background color is represented by column $C|VS+,Y$ in Table 3 and genetic variation is represented in number of SNPs compared to strain vaccinated against.
Influenza B and H3N2 consisted of the bulk of cases for the 2012-2013 flu season. Region 1 had the most cases for H3N2 with n = 126 and showed an average variation of 20.4 ± 4.2 SNPs (51% nonsynonymous), while variation was largest in Region 5 with 28.7 ± 7.4 SNPs (45% nonsynonymous). For influenza B, the most cases and most variation, excluding Region 7 that had just two cases, was seen in Region 4, with an average of 96.3 ± 73.6 SNPs (23% nonsynonymous) for n = 28 cases and smallest in Region 6 with n = 27 cases and an average variation of 50.7 ± 39.3 SNPs (25% nonsynonymous). Genetic variation from the H1N1 strain vaccinated against was largest in Region 2 with n = 5 cases and an average of 29.6 ± 0.9 SNPs (41% nonsynonymous) and was smallest in Region 3 with an average of 25.7 ± 1.2 SNPs (47% nonsynonymous) in n = 3 cases. During this flu season the overall amount of nonsynonymous mutations for H1N1, H3N2, and influenza B were 46%, 49%, and 24%, respectively.

**Figure 4.** 2013-2014 flu season by HHS Region, indicated in boldface. Background color is represented by column C|VS+,Y in Table 3 and genetic variation is represented in number of SNPs compared to strain vaccinated against.

H1N1 was the predominant strain in our 2013-2014 flu season dataset. The most cases were in Region 6, with n = 66 cases and an average of 29.8 ± 2.2 SNPs (44% nonsynonymous) from the strain vaccinated against. Variation peaked for H1N1 in Region 3 with n = 4 cases and an average of 30.8 ± 2.8 SNPs and was smallest in Region 4 with an average of 28.8 ± 2.1 SNPs in n = 42 cases. Only Regions 3, 4, and 6 had any H3N2 cases, with n = 1, n = 2, and n = 3, respectively and only Regions 4, 6, and 9 had any influenza B cases with n = 2, n = 2, and n = 5, respectively. The average variation in Region 9 for influenza B was, on average, 153.2 ± 58.8 SNPs, 21% of which were nonsynonymous. During this flu season the overall amount of nonsynonymous mutations for H1N1, H3N2, and influenza B were 45%, 53%, and 21%, respectively.

In addition to the HA gene, some clinical records also contained GenBank accession IDs for each gene in the genome of the influenza specimen that infected the patient. For the 2012-2013 flu season, 134 of the 470 records contained an accession ID for the neuraminidase (NA) gene, another surface protein on influenza viruses in addition to HA. Three
of these records were from H1N1 cases, 129 were from H3N2 cases, and two were from influenza B cases. We also compared the NA strains with the corresponding NA reference sequences: NC_026434.1, KJ942682, and CY115185 for H1N1, H3N2, and B, respectively. Overall, H1N1 showed an average of 14.0 ± 1.0 SNPs (40% nonsynonymous), H3N2 showed 11.0 ± 3.9 SNPs (26% nonsynonymous), and influenza B showed 28.5 ± 2.1 SNPs (39% nonsynonymous) from their respective reference sequences. We did not further analyze the NA gene by HHS Region due to the limited samples.

Discussion

Our analysis has implications for a number of topics that overlap from both the public and clinical health perspective. For example, understanding trends in vaccination efficacy based on strain, emerging drug resistance, and patient drug response given their infection type are a few of many possible applications. However, making such inferences remain difficult given the lack of integration between clinical and viral genetic data. As seen in Figure 2, the genetic variation between cases and the strains vaccinated against was highest in the 2012-2013 influenza B cases and the 2013-2014 influenza A/H3N2 cases; however only seven cases exist for the latter, making the former, with 132 cases, a far more interesting sample. In Figure 3, we can see that influenza B variation is high per HHS Region (Region 1 had no influenza B cases for that season), which is consistent with the overall trend in high variation. Genetic variation among 2012-2013 A/H3N2 and 2013-2014 A/H1N1 cases was quite small, surprising given the large number of cases per year for those strains (n = 313 and n = 225, respectively). Unfortunately, 52.3% of all cases in the former (Table 2) had an unknown vaccination status, making it difficult to match a trend to these data. Reporting of vaccination status was greatly improved between the 2012-2013 and 2013-2014 flu seasons, with the percentage of unknown cases falling from 40.4% to just 6% between the seasons.

This improved reporting of vaccination status made it possible to reliably compare the statuses with genetic variation for the 2013-2014 season. As seen in Figure 4, the H1N1 virus has variation in the northern regions than in the southern ones, while no type B cases are observed in the northern regions, although data was sparse for this and the H3N2 viruses (Figure 1). Unfortunately, it is difficult to compare trends between the two seasons we analyzed because the reported number of cases per virus per year decreased drastically for H3N2 and B but increased drastically for H1N1, as seen in Figure 1. Regardless, the visualization of these trends indicate that perhaps a bold step such as regionalization of vaccines to better capture and protect against representative virus strains in circulation may prove beneficial to millions of individuals. We will consider this for future work, perhaps culminated by a meta-analysis of vaccination efforts, as completed by Osterholm et al. 16. This could be combined with a viral genetic analysis like we have shown here.

There are several notable amino acid changes in the HA gene that have shown to enhance transmissibility of H1N1 viruses, including the simultaneous 222G and 163E mutations27 as well as an independent 219K mutation28. None of these mutations were found in the samples from either flu season analyzed. A genetic motif (RERRRKKR) that has been linked to HA cleavage and a potential target for drugs against H5N1 influenza29 was an available column in the clinical metadata file but was generally left blank. No sequence was found to contain this motif, perhaps an indication that it is conserved within H5 viral clades. The analysis of the NA gene was also critical because the main pharmaceutical treatment of influenza, oseltamivir, is a neuraminidase inhibitor and several mutations in this gene have been shown to cause oseltamivir resistance, including 119V, 292K, and 274Y30. None of these mutations were observed in any of the 129 NA sequences. According to the CDC, oseltamivir resistance was seen in 0.7% (n = 687), 0.2% (n = 2,440), and 0.0% (n = 1,044) of tested H1N1, H3N2, and influenza B samples tested during the 2012-2013 flu season31. It is possible that our lack of similar findings is simply a result of the drastically smaller sample size.

The authors recognize several limitations of this study. These include the lack of control data that could have been used to measure true efficacy and effectiveness of the vaccines rather than our C|VS+,Y representation. This could have led to more confident assessments of potential regionalization of vaccine production. It is unknown whether control cases were intentionally excluded from these data sources or there were simply no control cases to report (i.e. no negative tests for influenza at the reporting clinics). Furthermore, the clinical data we obtained and used via IRD was hardly complete and many columns were entirely devoid of data. Examples of columns that were generally blank
included transmission properties such as “Increased Virulence” and “Enhanced Transmission to Human”, and pre-
visit and post-visit medications prescribed to the patients. While we were able to gather useful information from the
columns where data was entered, it shows that we were only able to represent a small portion of the potential of these
data. Having information about virus properties, particularly that of resistance, could have enabled us to
geographically visualize these trends as well, similar to Figures 3 and 4, but with perhaps more significance due to the
ever present concerns over enhanced mutation and transmissibility. Other limitations include the aforementioned lack
of specific information regarding vaccine contents, forcing the use of a random HA sample for each of the strains
vaccinated against. Data regarding where the patients were vaccinated and the specific vaccine that they received
would have been invaluable for geographic accuracy and reliability of this study. Patient symptoms were listed for
nearly every case, but we did not explore possible ways to capture severity of the cases based on these symptoms.

An innovative combination of viral genetic data with clinical records would certainly fall under the umbrella of the
Precision Medicine Initiative32, especially in relation to resistance of oseltamivir and other antiviral drugs. If these
clinical records were outfitted with human genomic data as well as the viral sequences, it could highlight biomarkers,
genes, and SNPs associated with susceptibility to antiviral resistance. This would not strictly be limited to influenza
surveillance and could easily apply to concerning clinical bacterial infections like methicillin-resistant Staphylococcus
aureus (MRSA) or emerging zoonotic pathogens like Middle East respiratory syndrome coronavirus (MERS-CoV).
While this paper constituted a review of one application of linking clinical records with genetics of an infectious agent,
the potential for an innovative informatics pipeline to identify medication-resistant motifs and associate them with
clinical symptoms, human genomic markers, and geography is a challenge yet to be adequately addressed. In order
for this to occur, however, it appears that more clinics would need to be willing to begin outfitting their records with
links to pathogen DNA.

In this study, we demonstrate the value of integrating clinical data with viral genetics. While most clinical
microbiology laboratories do not have the financial or personnel resources to generate full genomic viral sequences
for each clinical case, our work highlights the usefulness of how even a sample of cases can provide valuable insight
into geographic disparities to seasonal epidemics and vaccination effectiveness. While our geographic representations
are not unlike FluView by the CDC33, we have shown that individual clinics or research labs could create such maps
in real time as sequences become available. Clinics should be interested in these types of reports because data
demonstrating a differentiated diffusion of strains with particular mutations that are indicated for antiviral resistance
could inform physicians of optimal treatments. Furthermore, these data could show expected trends in viral properties
among patients entering the health centers with acute respiratory illnesses, increasing efficacy of care and reducing
healthcare costs to both the individuals and clinics. While ultimately the integration of human genomic data into health
records remains the ideal gold standard, accomplishing an additional step like the incorporation of virulence could
prove to be a valuable resource for reducing morbidity and mortality and improving overall public health and
biosurveillance of infectious diseases.

References
1. Tumpey TM, Basler CF, Aguilar PV, et al. Characterization of the reconstructed 1918 Spanish influenza pandemic
3. WHO. Influenza (Seasonal) Fact Sheet. 2014 March 2014 [cited 2015 11 March]; Available from:
   http://www.who.int/mediacentre/factsheets/fs211/en/
   March 11]; Available from: http://www.cdc.gov/flu/fluavaxview/coverage-1314estimates.htm
6. Beyer WEP, McElhaney J, Smith DJ, Monto AS, Nguyen-Van-Tam JS, Osterhaus ADME. Cochrane re-arranged:
   Support for policies to vaccinate elderly people against influenza. Vaccine. 2013 12/5/;31(50):6030-3.


12. CDC. Key Facts about Influenza (Flu) & Flu Vaccine. 2013 9 September 2014 [cited 2015 7 March]; Available from: http://www.cdc.gov/flu/about/keyfacts.htm


33. CDC. FluView. 6 March 2015 [cited 2015 27 February]; Available from: http://www.cdc.gov/flu/weekly/
Exploration of Temporal ICD Coding Bias Related to Acute Diabetic Conditions

Mollie McKillop, M.P.H., Fernanda Polubriaginof, M.D., Chunhua Weng, Ph.D.
Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract
Electronic Health Records (EHRs) hold great promise for secondary data reuse but have been reported to contain severe biases. The temporal characteristics of coding biases remain unclear. This study used a survival analysis approach to reveal temporal bias trends for coding acute diabetic conditions among 268 diabetes patients. For glucose-controlled ketoacidosis patients we found it took an average of 7.5 months for the incorrect code to be removed, while for glucose-controlled hypoglycemic patients it took an average of 9 months. We also examined blood glucose lab values and performed a case review to confirm the validity of our findings. We discuss the implications of our findings and propose future work.

1.0 Introduction
Administrative data, namely ICD (International Classification of Diseases) codes, have been widely used to identify disease specific cohorts of patients\textsuperscript{1,2}. Such codes are often used in clinical and health services research because they are easy to obtain, have low associated costs, and can be aggregated to form large study samples. The accuracy of study results derived from administrative data depends on how well a particular coding scheme can correctly describe a disease cohort of interest. It is especially important to document coding biases as EHRs become more widely adopted and relied upon for large-scale data reuse. Increasingly, EHRs are used to document macroscopic human conditions, or phenotypes, automatically feeding data for secondary re-use purposes such as clinical research, quality improvement, and public health initiatives\textsuperscript{3}. Such uses require high-quality data, which are often lacking in the EHR. Coding bias is important to document and characterize for diabetes, which is an increasingly prevalent disease and is a major source of morbidity and mortality. It is a leading cause of blindness, end-stage renal disease and cardiovascular disease and is associated with high healthcare costs\textsuperscript{4}. Coding bias related to complications of diabetes, specifically ketoacidosis and hypoglycemia, are particularly important to understand because they are acute, life-threatening conditions that require hospitalization. Accurate results from studies using the EHR to phenotype these patients must be aware of any coding bias related to these conditions.

The validity of ICD-codes for identifying patient groups has been challenged many times before and for a variety of conditions\textsuperscript{5,6}. These studies have shown that ICD codes are biased because concept definitions for codes are incomplete or are unsatisfactory in granularity. Moreover, variability in coding behavior also leads to incorrect code assignment. Researchers have previously questioned the validity and generalizability of ICD codes as applied to diabetes\textsuperscript{7,8}. However, these studies have not examined coding bias among complications of diabetes and uncontrolled glucose levels. To date, we could only find two studies that examined coding bias associated with ketoacidosis. They are small in scale and focus on data captured before 2009\textsuperscript{9,10}. We seek to build upon this work by validating these results and examining bias among a larger, more diverse, more recently treated population of patients. We also seek to report coding bias related to hypoglycemia, which to our knowledge has not previously been documented. Recognizing that human phenotypes are time-dependent, we aim to describe the temporal bias associated with ketoacidosis and hypoglycemia ICD-9 codes. Previous research has neglected the dynamic nature of such phenotypes when examining coding bias. Temporal bias is important to understand for accurate phenotyping and the full-realization of purported EHR benefits.

This study uses acute complications of diabetes and uncontrolled glucose as examples to investigate the temporality of ICD coding bias. Specifically, we ask the question, “do patients, who are initially coded for ketoacidosis or hypoglycemia, remain coded as such, despite controlling their glucose levels?” In other words, we hypothesize that patients who initially receive an ICD-9 code assignment for either ketoacidosis or hypoglycemia continue to be assigned these codes, despite little clinical evidence that such code assignment is reasonable. We examine the extent of this bias over time by using survival analysis to determine the time it takes, on average, for patients to have the incorrect code removed from their personal EHR. We also examine this bias for different disease subgroups (Type 1 versus Type 2) and among glucose-controlled and uncontrolled patients. Finally, we hope to describe our method in enough detail to allow other researchers to replicate our findings and test the strength of our conclusions by examining temporal coding bias among other acute conditions. This study was performed in compliance with the.
World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was approved by the Columbia University Medical Center Institutional Review Board.

2.0 Methods
2.1 Data description and processing

For our study we utilized the Columbia University Medical Center Clinical Data Warehouse (CDW), which contains 24 years of data on 4.5 million patients\(^1\). Data were extracted in a multi-step process as depicted in Figure 1 to generate our cohort for analysis. First, relevant ICD-9 codes were identified using the Center for Medicare and Medicaid (CMS) ICD-9 Code Lookup\(^1\). The relevant codes are listed in Table 1. All instances of these codes from 2004 to October 2014 along with corresponding fake patient-identification numbers and real timestamps for each code were extracted. This resulted in 4844 unique patients, 2242 with ketoacidosis and 2602 with hypoglycemia. Along with the ICD-9 codes, all HgA1c lab values for these patients along with fake patient IDs and real timestamps were generated from the CDW. Lab values occurring after the initial coding of ketoacidosis or hypoglycemia were selected. Patients were then separated by ketoacidosis and hypoglycemia. After selecting only the lab values occurring after the initial coding for ketoacidosis, 1309 patients were retained; after selecting only the lab values occurring after the initial coding for hypoglycemia, 1129 patients remained. In the ketoacidosis subgroup, 112 patients were selected for analysis based on having controlled glucose levels, which was defined as having a median HgA1c lab value less than 7, a threshold recommended by the American Diabetes Association\(^1\). In the hypoglycemia group, 156 patients were selected for analysis based on having a similarly defined controlled glucose level.

<table>
<thead>
<tr>
<th>ICD-9 Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>250.10</td>
<td>DIABETES WITH KETOACIDOSIS, TYPE II OR UNSPECIFIED TYPE, NOT STATED AS UNCONTROLLED</td>
</tr>
<tr>
<td>250.11</td>
<td>DIABETES WITH KETOACIDOSIS, TYPE I [JUVENILE TYPE], NOT STATED AS UNCONTROLLED</td>
</tr>
<tr>
<td>250.12</td>
<td>DIABETES WITH KETOACIDOSIS, TYPE II OR UNSPECIFIED TYPE, UNCONTROLLED</td>
</tr>
<tr>
<td>250.13</td>
<td>DIABETES WITH KETOACIDOSIS, TYPE I [JUVENILE TYPE], UNCONTROLLED</td>
</tr>
<tr>
<td>251.10</td>
<td>OTHER SPECIFIED HYPOGLYCEMIA</td>
</tr>
<tr>
<td>251.20</td>
<td>HYPOGLYCEMIA UNSPECIFIED</td>
</tr>
</tbody>
</table>

2.2 Data analysis
Survival analysis was performed separately on the remaining 112 ketoacidosis and 156 hypoglycemia observations to determine the probability of remaining coded for ketoacidosis or hypoglycemia, despite having controlled glucose levels. Survival analysis was chosen as the statistical method for analysis because it allows for the time to an event to be computed. In this case, the event was decoding of a glucose-controlled patient. For the ketoacidosis group, the granularity of the ICD-9 codes allowed us to further explore any coding bias among disease subgroups.

First, individuals who were identified as Type 1 or Type 2 ketoacidosis patients by ICD-9 codes had their time to decoding compared and assessed for statistical significance using a cox proportional hazards test. Second, individuals who were identified as uncontrolled or controlled ketoacidosis patients by ICD-9 codes had their time to decoding compared and assessed for statistical significance again using a cox proportional hazard test. Because HgA1c lab tests measure a patient’s average glucose level over several weeks, we looked at the blood glucose lab tests 14 days before any ICD-9 code timestamp after the first ICD-9 code assignment. This time frame was chosen based on administrative timetables, since reimbursement claims are generally submitted within two weeks of any services or procedures provided. Also, the ICD-9 timestamp represents the discharge date. Considering the acute nature of the diseases under study, we felt this was an adequate time frame to select blood glucose lab results from.

The blood glucose test was selected because it provides a way to characterize patients as either in a state of hypoglycemia or ketoacidosis. According to UpToDate, a blood glucose serum level over 350 can be used to diagnose ketoacidosis\(^1\) while a blood glucose serum level under 40 can be used to diagnose hypoglycemia\(^1\).
blood glucose lab values were extracted from the CDW using the Medical Entities Dictionary, which is a large repository of medical concepts that are drawn from a variety of sources either developed or used at NYP. For each patient, the mean and median percentage of times these blood glucose lab values indicated disease for either the ketoacidosis or the hypoglycemia group was calculated. The analysis was performed using Rstudio version 0.98.1091. The datasets and R code are available for public use by contacting the primary author (MM).

Finally, a case review of deceased patient notes for both the ketoacidosis group and the hypoglycemia group was performed by a physician (FP) to further examine the validity of our results and perform an error analysis. Since patients were selected from groups dependent upon ICD-9 code assignment, the clinical reviewer provided a comprehensive chart review of the deceased patient notes from both the ketoacidosis and hypoglycemia group. The main purpose of this review was to contribute to an error analysis and provide a definitive “gold standard” of whether the patient had ketoacidosis or hypoglycemia. The Annals of Emergency Medicine guidelines for chart review were followed as much as possible. The participating reviewer (FP) is a clinician and knew the purpose of the study but not the study’s full outcome. The variables to be collected from the chart, as well as how these variables are defined, were determined prior to and documented for the coder. Variables were selected using UptoDate Guidelines to construct coding rules. Once a positive diagnosis based on the coded variables was determined by the reviewer, no further case review for that patient was performed. Overall, the reviewer examined 5 ketoacidosis patients and 13 hypoglycemic patients.

3.0 Results
3.1 Descriptive statistics
In total there were 112 patients in the ketoacidosis group and 156 patients in the hypoglycemia group. All patients that occurred in the ketoacidosis group also occurred in the hypoglycemia group. In the ketoacidosis group, there were a median of 5 HgA1c lab values per patient with a minimum of 1 and a maximum of 40. In the hypoglycemia group, there were a median of 10 HgA1c lab values per patient with a minimum of 2 and a maximum of 203. The trend in HgA1c lab values over time by patient for the ketoacidosis group (left) and hypoglycemia group (right) is shown in Figure 2.

![Figure 2](image-url)

**Figure 2.** HgA1c linear trend for ketoacidosis patients (left) and hypoglycemia patients (right). Each line represents a unique patient and colors are used to help distinguish different lines. Horizontal axis is sequential order of HgA1c lab tests; Vertical axis is HgA1c lab value.

A generalized estimating equation, assuming data were missing completely at random, was fit for the ketoacidosis and hypoglycemia groups separately to test for significant linear trend associations between HgA1c and time. The hypoglycemia group was found to have a significant association (p-value = 0.0045) while the ketoacidosis did not have such an association (p-value = 0.11). Both tests had a negative coefficient. In the ketoacidosis group, there were a median of 15.5 ICD-9 code assignments per patient with a minimum of 1 and a maximum of 211. In the hypoglycemia group, there were a median of 10 ICD-9 code assignments per patient with a minimum of 2 and a maximum of 203. Among the ketoacidosis group split by diabetes type as determined by ICD-9 codes, there were 39 Type 1 patients and 73 Type 2 patients. Among the ketoacidosis group split by uncontrolled versus controlled diabetes type as determined by ICD-9 codes and not median HgA1c lab values, there were 53 ‘not stated as uncontrolled’ patients and 59 ‘uncontrolled’ patients.

3.2 Survival analysis for the ketoacidosis group
The Kaplan-Meier survival curve for the 112 observations is shown in Figure 3. Time was recorded in 3-month intervals to reflect national guidelines for diabetes screening. According to the survival curve, it takes approximately 7.5 months for 50% of patients, who were initially coded for ketoacidosis and have their glucose in control, to stop being assigned any of the four ICD-9 codes listed in Table 1.

![Survival Curve for all Patients](image)

**Figure 3.** Kaplan-Meier survival curve for all ketoacidosis patients. Time is in 3-month intervals. Dashed lines represent confidence intervals for the survival curve.

A cox-proportional hazard test was performed to determine if significant differences existed between any of the four ICD-9 codes. This test was found to be significant at a level of .01 (p-value=0.00538), indicating at least one of the ICD-9 code survival curves is different from the others. This difference was examined further by separating the population into different groups based on diabetes type and 'uncontrolled' versus 'not stated as uncontrolled' ICD-9 code assignment. The survival curves for Type 1 and Type 2 diabetes are displayed in Figure 4. These curves were not significantly different from each other at a level of .05 (p-value=0.316) using the cox-proportional hazards test.

![Survival Curve of 'Type1' vs. 'Type2'](image)

**Figure 4.** Kaplan-Meier survival curve for ketoacidosis patients by diabetes type. Time is in 3-month intervals. The blue curve represents Type 1 diabetics and green curve represents Type 2 diabetics.

A third cox-proportional hazard test determined a significant difference (p-value=.00821) at a level of .01 between the Kaplan-Meier survival curves of patients coded for ‘uncontrolled’ (ICD-9 codes 250.12 and 250.13) versus
patients coded for ‘not stated as uncontrolled’ (ICD-9 codes 250.10 and 250.11). The survival curves are displayed in Figure 5 and indicate patients coded for ‘not stated as uncontrolled ketoacidosis’ (blue line) are decoded from any ketoacidosis ICD-9 code 25.1% faster than patients coded for ‘uncontrolled ketoacidosis’ (green line).

Figure 5. Kaplan-Meier survival curve for ketoacidosis patients by ICD-9 code assignment for ‘Not Stated as Uncontrolled’ (ICD-9 codes 250.10 and 250.11) vs. ‘Uncontrolled’ (ICD-9 codes 250.12 and 250.13). Time is in 3-month intervals. The blue curve is for ‘Not Stated as Uncontrolled’ diabetics and the green curve is for ‘Uncontrolled’ diabetics.

3.3 Glucose lab value assessment for the ketoacidosis group

Patients were selected from those that had one or more ICD-9 code assignments for ketoacidosis after the initial coding for the disease. The glucose lab tests occurring 14 days before any ICD-9 code timestamp, except the first timestamp, were selected and assessed for positive indication of ketoacidosis based on a blood glucose cutoff level of 350 mg/dL as determined by UpToDate\(^4\). Only 90 out of the 112 patients had blood glucose labs occurring 14 days before the ICD code timestamp values under consideration. Positive disease indications were assigned a value of one and tabulated by patient. The mean number of positive disease indications was 210 with a median of 195, while the median percentage of positive disease indications was 6.33% of all blood glucose lab tests per patient. The mean percentage of positive disease indications over the number of blood glucose labs was calculated per patient and the distribution is shown in Figure 6.

Figure 6. Distribution of mean positive ketoacidosis indication by sequential ICD code assignments.
3.4 Survival analysis for the hypoglycemia group
The Kaplan-Meier survival curve for the 156 hypoglycemia patients is shown in Figure 7. Time was recorded in 3-month intervals to reflect national guidelines for diabetes screening. According to the survival curve, it takes 9 months for 50% of patients, who were initially coded for hypoglycemia and have their glucose in control, to stop being assigned any of the ICD-9 codes for ketoacidosis. Unfortunately, no further survival analysis of hypoglycemia subgroups could be performed due to the limited granularity of ICD-9 codes in this category.

![Kaplan-Meier survival curve](image)

**Figure 7.** Kaplan-Meier survival curve for all hypoglycemia patients. Time is in 3-month intervals. Dashed lines represent confidence intervals for the survival curve.

3.5 Glucose lab value analysis for the hypoglycemic group
Patients were selected from those that had one or more ICD-9 code assignments for hypoglycemia after the initial coding for the disease. The glucose lab tests occurring 14 days before any ICD-9 code timestamp, except this first timestamp, were selected and assessed for positive indication of hypoglycemia based on a blood glucose cutoff level of 40 mg/dL as determined by UpToDate. Only 33 out of the 156 patients had blood glucose labs occurring 14 days before the ICD code timestamp values under consideration. Positive indications were assigned a value of one and tabulated by patient. The mean number of positive disease indications was 120 with a median of 109, while the mean percentage was .8977% and the median percentage was 0%. The percentage of positive disease indications over the number of blood glucose labs was calculated per patient and the distribution is shown in Figure 8.
4.0 Discussion

4.1 Major Implications

The results of this study seem to suggest a coding bias related to ICD-9 ketoacidosis and hypoglycemia codes. It appears that those who are initially coded for ketoacidosis or hypoglycemia continue to be coded as such despite having their glucose levels in control, as indicated by median HgA1c lab values. This finding is also supported by the low median percentage of blood glucose lab tests indicating disease for the ketoacidosis group (6.33%) and the hypoglycemia group (0%). More importantly, this study examines the temporal nature of such coding bias. Because ketoacidosis and hypoglycemia are conditions that last several hours to days, our results show the extent of this bias seems severe when the temporality of coding behavior is considered. This bias persists when diabetes type and control, as determined by ICD-9 codes, are examined. For example, patients that are ICD-9 coded for ‘uncontrolled diabetes, ketoacidosis’ and have controlled glucose levels (as determined by HgA1c lab tests) have their inaccurate code removed 79.9% slower than the same patients coded for ‘not stated as uncontrolled diabetes, ketoacidosis’. This makes clinical sense considering those that have uncontrolled diabetes have a higher probability for ketoacidosis, a complication related to poor glucose control. With regard to hypoglycemia, temporal coding bias seems to remain and the effect is even stronger than in the ketoacidosis group. This may be due to differing disease severity in that hypoglycemia is a less severe disease than ketoacidosis, so more people are likely to be coded for its presence. However, it remains unclear why patients, who have lab values indicating their glucose is already controlled, continued to receive hypoglycemia codes. We hypothesize that because hypoglycemia is a less severe and a less formally defined disease than ketoacidosis, it is easier to interpret this diagnosis from patient notes and lab values. This allows for coders to more frequently assign hypoglycemia ICD-9 codes. Our findings are similar to previous studies that found ICD-9 coding bias among ketoacidosis patients. Moreover our results imply a potential overestimation of disease prevalence if ICD-9 codes are used solely to identify a ketoacidosis or hypoglycemia cohort.

4.2 Error analysis

In the ketoacidosis group all 5 patients were positively identified by manual case review as having ketoacidosis. In the hypoglycemia group, 5 patients did not have notes at any point two weeks before the first ICD code assignment, leaving 8 code assignments for our error analysis. It remains unknown as to why these notes are missing. Of these 8 assignments, 7 were confirmed as correct (patients had hypoglycemia) with the remaining incorrect code due to a patient having normal glucose levels. Major sources of error therefore include patients with normal glucose being coded for hypoglycemia as well as no clinical basis for code assignment, since no notes existed for several patients.

In addition to the error analysis, we also identified deceased patients who had repeated ICD-9 code assignments for their particular disease group. The purpose of this assessment was to see if repeated ICD-9 code assignments were
valid. Only one patient fit this description in the ketoacidosis group. In a case review performed in the same manner as the error analysis, the patient was identified as having normal glucose. In the hypoglycemia group, all 13 patients had repeated ICD-9 code assignments, with a total of 16 repeated assignments (one patient had 4 repeated assignments after the first assignment, while the rest of the patients had one). This analysis revealed that 8 of these ICD code assignments were missing all notes two weeks before the code assignment. Of the 8 remaining code assignments, 6 of these were found to be correct through case review of patients’ notes. The other two assignments for hypoglycemia were assigned to patients with normal glucose.

The survival analysis and mean blood glucose results along with the case review suggest a semantic gap between the physicians who author such notes and the coders who ultimately assign ICD-9 codes. The physician’s concept of the patient must be encoded in the patient notes and then from this computable representation the coder must form a concept of an ICD-9 code. In this process it appears that clinical information is lost or misunderstood. Either the concept is not encoded (i.e. no notes are present two weeks before a code assignment) or the concept is semantically altered once present in the EHR. The coding bias we describe could be partially explained by the over-use of copy-pasted text, which appeared multiple times in each of the case-reviewed patient notes. Finally, economic incentives for incorrectly coding patients should also be considered and further studied as a potential source of coding bias.

4.3 Limitations

This study has several limitations. First, the sample size of 112 ketoacidosis patients and 156 hypoglycemia patients is relatively small, and larger studies could reveal interesting differences in ketoacidosis coding between Type 1 and Type 2 patients. Moreover a larger dataset could allow for more complex measures of glucose control. Because the median number of HgA1c lab tests performed was only 5 for the ketoacidosis group, more sophisticated approaches, such as time series analysis, that can measure trends in glycemic control were unfeasible. Moreover, our goal was not to model HgA1c trends but to characterize a group of patients with reasonably well-controlled glucose levels. Therefore, median HgA1c lab values were used as a measure for glucose management. However, we feel this is a reasonable assessment of glucose control in relation to the acute complications we are exploring. Finally, it should be noted that all HgA1c measurements were collected observationally, and that there may be high variability in measurement quality.

In addition, it is possible our findings are due to repeated hospitalizations for patients who do in fact have either hypoglycemia or ketoacidosis, yet this seems unlikely given the low median percentage of glucose blood tests indicating either disease and the follow up we performed in our error analysis. Finally, the two groups of patients are not independent from each other, in that all ketoacidosis patients are also represented in the hypoglycemia group. This limits the generalizability of our findings to other patient groups.

Importantly, there may be outliers influencing the results of our analysis, as is implied by the relatively large difference between the number of maximum ICD-9 code assignments and median number of ICD-9 code assignments for both the ketoacidosis group and the hypoglycemia group. We did do a visual evaluation to identify these points, plotting the patient id against the number of ICD code assignments for both disease groups. There appeared to be two outliers among the ketoacidosis group and another two outliers in the hypoglycemia group. We did not remove these outliers since our aim was exploratory in nature and there appeared to be few outliers. In addition, we do not wish to propose a mathematical model for temporal coding bias but rather to prove the feasibility of a method for exploring this phenomenon. More robust and larger datasets could aid in determining why such outliers might exist.

Only 18 deceased patients’ records (5 in the ketoacidosis group and 13 in the hypoglycemia group) were reviewed. These patients may not represent each disease group accurately. Because only deceased patients’ notes were used for the error analysis, bias may be introduced by selecting patients who are perhaps more likely to die, and therefore more sick, than other patients. However, we feel this bias is limited, since its presence would shift our results toward concluding that there is less coding bias (as it is more likely that coders would assign ICD-9 ketoacidosis and hypoglycemia diagnosis codes). On the other hand, these patients may be so sick that coders fail to identify ketoacidosis and hypoglycemia as the patient’s primary reason for hospital visit in favor of codes representing more serious diagnoses. Yet, ketoacidosis and hypoglycemia are very serious life-threatening conditions that would probably outrank most other diseases in terms of prominence and importance during a hospital visit.

Finally, the numeric results of this study are probably not generalizable to other settings, because decoding of ketoacidosis could be due to either the coder stopping ketoacidosis code assignment or, perhaps more likely, the patient not returning to the New York-Presbyterian Hospital for care. In either case, the extent of this bias would alter our results in the direction of less temporal coding bias.
4.4 Future work

Future work should try to examine the reasons for the documentation discrepancies between physicians and coders as well as any financial motivations coders might have for assigning incorrect codes. Physicians could try to be more consistent in their terminology and use less copy-pasted text, so that coders have a clearer representation of the ketoacidosis and hypoglycemia concept. Coders and physicians could be better educated on the importance of communicating and picking accurate ICD-9 codes. In addition, more research should be done to develop accurate and automated methods of ICD-9 code assignment, which could overcome the coding discrepancies between physician and coder. This study also reveals challenges in secondary reuse of clinical data for research purposes. More relevant organization of clinical data would allow for easier data analysis and increase research efficiency. For example, this could occur through front-end applications that allow users to view, sort and analyze clinical data.

Conclusion

The results of this study indicate that temporal coding bias is a problem among patients with acute complications related to poor glucose control. This bias is significantly different between those coded as ‘not stated as uncontrolled’ vs. ‘uncontrolled’ with ICD-9 ketoacidosis codes and occurs among two different acute complications, i.e., ketoacidosis and hypoglycemia. A case review of patients’ notes further confirmed this bias. We also contribute a novel method for coding bias research related to acute diseases. Understating such bias is increasingly important as the EHR becomes more widely used for phenotyping and cohort selection. Future work should have a socio-technical approach, evaluating how these biases arise. Specifically, the coding discrepancies between physicians and coders should be further explored and documented.

Acknowledgements

This study was sponsored by the U.S. National Library of Medicine grants R01LM009886 (PI: Weng) and T15LM007079 (PI: Hripcsak) and U.S. National Center for Advancing Translational Science grant UL1 TR000040 (PI: Ginsberg).

References


Classification of Clinically Useful Sentences in MEDLINE

Mohammad Amin Morid, MS; Siddhartha Jonnalagadda, PhD; Marcelo Fiszman, MD, PhD; Kalpana Raja, PhD; Guilherme Del Fiol, MD, PhD

a Department of Operations and Information Systems, David Eccles School of Business, University of Utah, Salt Lake City, UT, USA
b Department of Preventive Medicine, Division of Health and Biomedical Informatics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
c Lister Hill Center, National Library of Medicine, Bethesda, MD, USA
d Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

Abstract

Objective: In a previous study, we investigated a sentence classification model that uses semantic features to extract clinically useful sentences from UpToDate, a synthesized clinical evidence resource. In the present study, we assess the generalizability of the sentence classifier to Medline abstracts.

Methods: We applied the classification model to an independent gold standard of high quality clinical studies from Medline. Then, the classifier trained on UpToDate sentences was optimized by re-retraining the classifier with Medline abstracts and adding a sentence location feature.

Results: The previous classifier yielded an F-measure of 58% on Medline versus 67% on UpToDate. Re-training the classifier on Medline improved F-measure to 68%; and to 76% (p<0.01) after adding the sentence location feature.

Conclusions: The classifier’s model and input features generalized to Medline abstracts, but the classifier needed to be retrained on Medline to achieve equivalent performance. Sentence location provided additional contribution to the overall classification performance.

Introduction

Most clinical questions raised by clinicians in the course of patient care can be answered by online primary literature resources, such as Medline. However, there are critical barriers to the use of the primarily literature at the point of care. Specifically, clinicians need to search, screen, appraise, and integrate literature findings into their decision making for a particular patient. This process is labor intensive and not compatible with busy clinical workflows. Several approaches have been pursued to improve efficient consumption of the primary literature, including improvements in the search process, question and answering systems, and automatic summarization of Medline abstracts and full-text articles. Despite substantial progress in these approaches, recent studies still show that clinicians prefer distilled recommendations in the form of guidelines and evidence summaries.

Significant effort has been dedicated to automatic biomedical text summarization. Yet, most previous studies aimed at generating summaries that resemble article abstracts written by study authors. However, article abstracts are written to summarize all elements of a study, such as purpose, methods, results, and conclusions. On the other hand, for patient care decision making, clinicians prefer sentences that provide patient-specific, actionable recommendations for a particular intervention as opposed to general background and study methods. For example, the sentence “Apixaban 2.5 mg twice daily, starting on the morning after total knee replacement, offers a convenient and more effective orally administered alternative to 40 mg per day enoxaparin, without increased bleeding” provides an actionable treatment finding for patients who undergo total knee replacement. Specific methods are needed for extracting clinically useful sentences from clinical studies.

In a previous study we developed a feature-rich classification model for extracting clinically useful sentences from synthesized evidence resources, such as UpToDate. The study was based on 4,824 sentences from 18 UpToDate documents on the treatment of six chronic conditions: coronary artery disease, hypertension, depression, heart failure, diabetes mellitus, and prostate cancer. In the present study, we attempt to apply the sentence classifier to the primary literature. Specifically, the study has two main goals: (1) to assess the generalizability of the feature-rich
classifier on extracting clinically actionable statements from PubMed abstracts; and (2) to assess if optimization of
the classifier for PubMed abstracts results in improved classification accuracy.

Background and Significance

In a previous study we designed and assessed a method for extracting clinically useful sentences from synthesized
online clinical resources. The method’s underlying assumption is that clinically useful sentences are actionable
statements that provide a specific recommendation for an intervention (e.g., medication treatment) that should be
employed with a specific patient population. To capture these characteristics, the method uses three sets of semantic
features from the PubMed abstracts. The method consists of a Kernel-based Bayes Network classification model
with Gaussian kernel density estimators that classifies each sentence as clinically useful or not. As shown in
previous research, the Kernel-based Bayesian Network is robust to highly imbalanced datasets such as the one used
in this paper. This classifier is a Bayesian Network that estimates the true density of the continuous variables
using kernels, which are weighting functions used to estimate random variables’ density function. The classifier is
based on three domain-specific feature types extracted from UpToDate sentences: 1) treatment-related UMLS
concepts and their semantic groups; 2) semantic predications; and 3) patient population. A summary of these
features is provided below.

The first set of features consists of treatment-related UMLS concepts, and their corresponding semantic groups,
extracted from sentences using MedTagger, which is an extension of the cTAKES natural language processing
(NLP) pipeline. The UMLS semantic group of each concept was obtained, leading to four features according to the
following semantic groups: Chemicals & Drugs (CHEM), procedures (PROC), physiology (PHYS), and disorders
(DISO).

Semantic predications are relations that consist of a subject, a predicate, and an object. The sentence classifier uses
treatment-related predications extracted by SemRep, a semantic NLP parser that uses underspecified syntactic
analysis and structured domain knowledge from the UMLS. Six types of predications were extracted as features: TREATS/NEG_TREATS,
ADMINISTERED_TO /NEG_ADMINISTERED_TO, AFFECTS/NEG_AFFECTS, PROCESS_OF /
NEG_PROCESS_OF, PREVENTS / NEG_PREVENTS, and COMPARED_WITH / HIGHER_THAN /
LOWER_THAN / SAME_A. For instance, from the sentence below:

> Adding corticosteroid injection to conventional treatment in hemiplegic shoulder pain improved shoulder range of
motion and decreased pain scores before treatment to the first and fourth weeks of treatment.

SemRep produces the following output:

  Shoulder Pain PROCESS_OF Hemiplegics
  Injection procedure TREATS Shoulder Pain
  Adrenal Cortex Hormones TREATS Shoulder Pain

which yields the following features:

  Total number of predications: 3
  PROCESS_OF instances: 1
  TREATS instances: 2

Finally, patient population determines whether a sentence includes a description of the types of patients who are
eligible to receive a certain treatment based on a pattern-based method. This produced one binary feature, which
indicates whether a sentence describes the target population or not. The method uses the Stanford lexical parser
and Tregex. The Tregex patterns are similar to regular expressions, but more advanced in extracting patterns such
as a noun phrase with two consecutive prepositional phrases, a verb phrase with two consecutive prepositional
phrases, and a noun phrase preceding a subordinating conjunction. For example, in the following sentence the
population extraction algorithm identifies that the sentence includes a target population (“patients with advanced
NSCLC),

> The addition of vandetanib to docetaxel provides a significant improvement in PFS in patients with advanced
NSCLC after progression following first-line therapy.”
In the present study, we test the generalizability of the described feature-rich sentence classifier to the primary literature and whether optimization of the sentence classifier results in performance gains.

**Methods**

The study methods consisted of the following steps: 1) development of a gold standard of clinically useful sentences from PubMed abstracts; 2) extraction of the three feature categories (i.e., concept, predication and population features) for sentence classification; 3) optimization of the sentence classifier to identify clinically useful sentences from PubMed abstracts; and 4) assessment of classifier performance.

All features for our sentence feature-rich sentence classifier in different experiments are summarized in Table 1. Specifically, these are the inputs for the Kernel-based Bayes Network classification model with Gaussian kernel.

**Gold Standard.** The gold standard consisted of 2,146 sentences from 140 PubMed abstracts that were randomly selected from 34,913 PubMed citations of high quality clinical studies published between January 2010 and October 2014. We focused on high quality clinical studies because they are likely to be more useful for patient care decision making. Citation quality was determined using the classifier developed by Kilicoglu et al.\(^4\) Sentences from the selected citations were retrieved from the SemanticMedline database, which contains sentences, and their semantic predications, extracted from all abstracts in Medline\(^25\).

For structured PubMed abstracts, we found that the gold standard contained clinically useful sentences only in the conclusion and results sections. Thus, we excluded all the sentences that were not in these sections. This filtering was done using the NlmCategory tag of the Medline citations in XML format, which provides standard section categories (e.g., METHODS, RESULTS, CONCLUSIONS) for abstracts that are written in a structured format. For unstructured abstracts, we included all sentences. As a result, the dataset was narrowed to 954 sentences from 124 structured abstracts and all 118 sentences from 16 unstructured abstracts (i.e., total of 1,072 sentences).

Next, the sentences in the gold standard were rated by one of the study authors (GDF) according to a validated clinical usefulness scale (Table 2), which was slightly adapted from one of our previous studies\(^26\). Sentences are rated from 1 to 4, with 4 being the most useful. The core principle of this scale is that clinically useful sentences follow the PICO format, i.e. sentences that define the study patient population, the intervention under investigation, the comparison (e.g., placebo), and the study outcome. The PICO format has been recommended to clinicians for formulating well-structured clinical questions and has been applied in several biomedical information retrieval studies\(^8,27,29\).

**Table 1: Features used to develop the classification model.**

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Number of features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predication</td>
<td>7</td>
<td>Total number of predications with a treatment-related predicate (1 feature) and number of predication instances per treatment-related predicate (6 features).</td>
</tr>
<tr>
<td>Population</td>
<td>1</td>
<td>Whether or not a sentence includes a description of the types of patients who are eligible to receive a certain treatment.</td>
</tr>
<tr>
<td>Concept</td>
<td>5</td>
<td>Total number of concepts in the sentence (1 feature) and number of concept instances per UMLS treatment-related semantic group (4 features).</td>
</tr>
<tr>
<td>Location</td>
<td>1</td>
<td>Location of the sentences in the abstract, which can be either Conclusion, Results, or Unknown (unstructured abstracts)</td>
</tr>
</tbody>
</table>
Table 2: Clinical usefulness rating criteria.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sentences that, in isolation, don’t convey clear meaning.</td>
<td>“Lorazepam rescue was permitted after dose two.”</td>
</tr>
<tr>
<td>2</td>
<td>Background information, such as the epidemiology and physiopathology of a condition, mechanism of action of an intervention (e.g., a drug), justification for conducting the study, study objectives, and description of the study design (e.g., randomized controlled trial, systematic review).</td>
<td>“This phase III, randomised, double-blind, placebo-controlled, parallel-group study enrolled 344 individuals who received one, two or three doses of inhaled losapine (5 or 10 mg) or a placebo.”</td>
</tr>
<tr>
<td>3</td>
<td>Study findings without a population, comparison, intervention and outcome (PICO); or secondary study findings.</td>
<td>“Death or myocardial infarction rates were reduced by fondaparinux in tertile I (age&lt;56 years, 4.5% vs 4.8%, hazard ratio [HR] 0.94, 95% CI 0.71-1.25), in tertile II (age 56-68 years, 7.9% vs 9.7%, HR 0.80, 0.65-0.98), and in tertile III (age&gt;=69 years, 17.2% vs 19.8%, HR 0.87, 95% CI 0.75-1.01, P for heterogeneity=0.87).”</td>
</tr>
<tr>
<td>4</td>
<td>Primary study findings or treatment safety findings with a population [P], intervention [I], comparison [C], and outcome [O].</td>
<td>After adjustment for covariates, infants with CNS involvement [P] who had been randomly assigned to acyclovir suppression [I] had significantly higher mean Bayley mental-development scores at 12 months [O] than did infants randomly assigned to placebo [C] (88.24 vs. 68.12, P=0.046)., 4, 1, 1</td>
</tr>
</tbody>
</table>

The final dataset is available online for the research community. The distribution of sentences according to their ratings is shown in Table 3.

**Optimization strategies.** To optimize the feature rich classifier based on PubMed abstracts, two strategies were employed. First, the feature-rich classifier was re-trained on PubMed abstracts (instead of UpToDate documents) using the same features identified in our previous study. Second, sentence location was included as an additional feature to the sentence classifier. The location feature was extracted from structured abstracts using the NlmCategory tag of Medline citations in XML format. The possible values for the location feature were Conclusions or Results for structured abstracts, and Unknown, for unstructured abstracts.

**Assessment of classification performance.** We conducted three experiments to test the following hypotheses:

Hypothesis 1: The feature-rich classifier trained on UpToDate sentences has comparable performance on the primary literature. To test this hypothesis, we compared the performance of the sentence classifier when applied to the original UpToDate dataset versus Medline sentences. The goal was to assess the generalizability of the sentence classifier to the primary literature.

1 https://drive.google.com/file/d/0B08sY2K1TQg9OX0pIOHVzLTVTaTQ

2018
Table 3: Sentence distribution according to sentence usefulness ratings.

<table>
<thead>
<tr>
<th>Type</th>
<th>Rating</th>
<th>Total number of sentences</th>
<th>Average number of sentences per abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Useful</td>
<td>1</td>
<td>102 (10%)</td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>117 (11%)</td>
<td>2.60</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>750 (70%)</td>
<td>5.43</td>
</tr>
<tr>
<td>Useful</td>
<td>4</td>
<td>103 (10%)</td>
<td>1.12</td>
</tr>
</tbody>
</table>

Hypothesis 2: Re-training the feature-rich classifier on the primary literature improves performance compared to the original classifier. To test this hypothesis, we assessed the performance of the feature-rich classifier trained on Medline sentences compared to the original classifier, which was trained on UpToDate sentences. Also, the enriched feature-rich classifier was compared to a baseline classifier where all sentences in the Conclusion section of structured abstracts and the last 10% of the sentences in unstructured abstracts were labeled as clinically useful (i.e., positive class).

Hypothesis 3: Adding sentence location to the feature-rich classifier improves its performance on the primary literature. To test this hypothesis, we compared the performance of the re-trained feature-rich classifier enriched with a sentence location feature versus the re-trained classifier without sentence location.

Experiment procedures. Ordinal ratings were converted into binary values: sentences rated as “4” were considered as the positive class (i.e., clinically useful sentences) and the remaining sentences were considered as the negative class. As a result, 89% of the sentences in the gold standard were labeled as positive. This distribution is similar to the sentences in the UpToDate dataset, with 87% positive sentences.

For the first hypothesis the feature-rich classifier was trained on 4,824 UpToDate sentences from our previous study, and then tested on 1,072 Medline sentences. For the second and third hypotheses we employed a 20-fold cross-validation strategy with each fold containing 7 abstracts.

Finally, classification performance was measured according to the average precision, recall, and F-measure across the 20 folds. F-measure was defined a priori as the primary outcome for hypotheses testing. For statistical significance test of all experiments, first we applied the Friedman’s test to verify differences among multiple classifiers. If significant at an alpha of 0.05, pairwise comparisons were made with the Wilcoxon Signed-Rank test. This statistical approach is aligned with the method recommended by Demsar.

Results

Similar to UpToDate sentences, descriptive statistics of the sentences and features in the Medline gold standard show that all feature types were correlated with useful sentences.

Hypothesis #1: The feature-rich classifier trained on UpToDate sentences has comparable performance on the primary literature. The F-measure for the feature-rich sentence classifier on the Medline dataset was 58% versus 67% on UpToDate (p<0.01) (Figure 1).

Hypothesis #2: Re-training the feature-rich classifier on the primary literature improves performance compared to the original classifier. The re-trained feature-rich classifier performed significantly better than the original classifier on Medline sentences and the baseline (F-measure = 68% versus 58% and 45% respectively; p<0.001 for both comparisons) (Figure 2). Moreover, the performance of the re-trained feature-rich classifier on Medline abstracts was comparable to the performance of the feature-rich classifier on UpToDate sentences (F-measure = 68% versus 67%; p=0.53).
Hypothesis #3: Adding sentence location to the feature-rich classifier improves its performance on the primary literature. As seen in Figure 3, adding the location feature further improved the classifier performance (F-measure = 76% versus 68%, p<0.01).

Discussion

This study investigated an automated method for extracting clinically useful sentences from primary literature resources such as Medline. To achieve this goal, we employed and adapted a feature-rich sentence classification model developed in a previous study. Such a method can be used in clinical decision support tools that use automatic summarization to help clinicians integrate findings from the primary literature into their decision making routine. We are currently integrating the optimized sentence classifier into one of these tools, known as the Clinical Knowledge Summary (CKS)\textsuperscript{31, 32}. The CKS automatically summarizes patient-specific evidence from multiple resources and can be integrated with electronic health record (EHR) systems through the Health Level Seven (HL7) Context-Aware Knowledge Retrieval (Infobutton) Standard\textsuperscript{33, 34}.

We conducted three experiments to test different hypotheses. The first experiment showed that the classifier, trained on UpToDate sentences, loses accuracy when applied to Medline sentences. Specifically, the classifier’s precision significantly decreased on Medline compared to UpToDate, although its performance in terms of recall was equivalent. A possible reason is that Medline sentences have different syntactic and semantic structure from UpToDate sentences. UpToDate provides recommendations based on synthesis of the evidence provided by multiple studies (e.g., “In patients resistant to initial therapy with hydroxychloroquine (HCQ) or sulfasalazine (SSZ), we suggest adding methotrexate (MTX) or treating with a combination of HCQ, SSZ, and MTX, rather than switching to a TNF inhibitor or to a TNF inhibitor plus MTX.”). Original studies provide a conclusion of the study findings, but in most cases there is no clear recommendation for clinical practice (e.g., “In this treatment-refractory population, tofacitinib with methotrexate had rapid and clinically meaningful improvements in signs and symptoms of rheumatoid arthritis and physical function over 6 months with manageable safety.”)

In the second experiment, re-training the classifier on Medline sentences with the exact same features resulted in improved performance, equivalent to performance on UpToDate sentences. This finding
Figure 2: Average precision, recall and F-measure of the baseline method compared with the feature-rich sentence classifier in different training and testing settings.

Figure 3: Average precision, recall and F-measure of the feature-rich sentence classifier, with and without location feature, and trained and tested on Medline sentences.
confirms that the classifier’s model and features used for UpToDate are generalizable to Medline. Also, the retrained classifier outperformed a baseline classifier, which was just based on sentence location. This shows that advanced classification methods based on NLP techniques and machine learning algorithms are worth the gained performance and classification power. The last experiment confirmed the hypothesis that sentence location in Medline abstracts further improves classification performance. This finding was expected, since study authors often summarize the main study findings and their clinical implications in the conclusion section of Medline abstracts.

Analysis of false-positives and false-negatives showed two main error categories. The first category includes recommendations that were too general, such as in “Drug therapy is recommended to stabilize and relieve symptoms in patients with preserved ventricular function.” Future studies can try to address this issue by identifying general treatment concepts using UMLS concept hierarchies. The second category was clinically useful sentences for which SemRep and MedTagger were unable to extract predications and concepts, such as in “Augment™ may represent a safe and efficacious treatment alternative to ABG during foot and ankle arthrodesis.” Fine tuning of NLP methods are needed to address this kind of problem.

Limitations. The main limitation of this study is the use of Medline abstracts as opposed to full-text articles. Medline abstracts do not report all the conclusions of a study, therefore sentence classification is limited to clinically useful sentences available in the abstract. Moreover, the gold standard consisted of high quality clinical studies published in high impact journals, which have a higher rate of structured abstracts than other studies in Medline. Since the sentence classifier benefits from standardized abstract structure, the performance of the optimized classifier applied to a dataset with a higher rate of unstructured abstracts is likely to be lower.

Future studies. We are integrating the feature-rich sentence classifier with an interactive clinical decision support tool that provides patient-specific summaries of clinical evidence from UpToDate and Medline.3 Future studies also include applying and adapting the sentence classification method to full-text articles.

Conclusion

We investigated the generalizability of a feature-rich sentence classification model, which was trained on UpToDate sentences, to Medline abstracts. The feature-rich classifier’s model and input features were generalizable to sentences from Medline abstracts, but the classifier had to be retrained on those sentences to achieve equivalent performance. Optimization of the classifier by adding a sentence location feature improved classification performance. The resulting sentence classifier can be used as a component of text summarization systems to help clinicians’ patient care decision-making.

Acknowledgement

This project was supported by grants 1R01LM011416-01 and 4R00LM011389-02 from the National Library of Medicine.

References


Desiderata for Major Eligibility Criteria in Breast Cancer Clinical Trials

Matthew L. Paulson, MPH, Chunhua Weng, PhD
Department of Biomedical Informatics, Columbia University, New York, NY

Abstract
Use of major eligibility criteria is a popular but unstudied folk practice for improving patient screening efficiency for clinical studies. This mixed-methods research study derived the desiderata for major eligibility criteria in breast cancer clinical trials. We randomly selected thirty interventional breast cancer clinical trials conducted at The New York-Presbyterian Hospital on the Columbia University Medical Center campus to create training (N=20) and testing (N=10) datasets. We utilized the Think-aloud protocol to gauge how clinical researchers identify and use major eligibility criteria to prescreen patients for clinical trials during an audio-recorded interview. A focus group session was held to understand the current prescreening process and investigate how it could be optimized to maximize recruitment rates. Using the grounded theory method, we annotated transcriptions to discover user rationale and desiderata behind major eligibility criteria in breast cancer clinical trials, which were later evaluated in a follow-up survey.

Introduction
Recruitment to clinical trials remains the biggest barrier to clinical and translational research. Most clinical trial studies have dozens of complex inclusion and exclusion eligibility criteria. Therefore, when using these eligibility criteria for subject screening, a common and cost-effective practice adopted by clinical researchers is to start with a small set of “major eligibility criteria”. If a patient does not satisfy any major criterion, there will be no need to waste the time to go over the complete eligibility criteria list. Moreover, many clinical research coordinators know that a lot of minor eligibility criteria are either irrelevant or rare among the target population so that the information gained from those minor criteria returns minimal value for screening eligible patients. Therefore, they focus on using major eligibility criteria to maximize information gain during prescreening processes.

Understanding the definition and characteristics of major eligibility criteria has three foreseeable benefits: (1) facilitating automated selection of major eligibility criteria for efficient electronic clinical trial prescreening using electronic patient data; (2) enabling cost-effective analysis of existing eligibility criteria; and (3) enabling understanding of the prioritization behaviors of clinical researchers to inform the design of advanced clinical trial screening methods. However, to the best of our knowledge, little is known about the selection of major eligibility criteria for clinical trial prescreening. With a focus on the important disease area of cancer, in this study our research questions are “is there a common definition for major cancer clinical trial eligibility criteria?” and “if yes, what are the characteristics of these major eligibility criteria?” In this study, we define a desiderata for major eligibility criteria of clinical trials as well as provide examples of standardized major eligibility criteria using a mixed-methods approach using qualitative and quantitative measures with clinical researchers at Columbia University Medical Center. The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and was approved by the Columbia University Medical Center Institutional Review Board.

Background
1. Barriers to recruitment and the need for prescreening
Slow clinical trial recruitment is a significant obstacle to medical discovery. Recruitment obstacles can lead to trial completion delays from one to six months for most clinical trials. Automating a significant portion of the patient screening process could reduce recruitment delays by identifying potentially eligible patients. The prescreening of patients for clinical trials requires use of study specific predefined inclusion and exclusion eligibility criteria. The lack of standardization for eligibility criteria creates obstacles to comparing trials to one another. As a result, patient screening is a complex multi-step and multi-user process that requires significant and expensive human time and effort. Moreover, the complexity of eligibility criteria can cause potentially eligible patients to be overlooked. Information systems have the potential to decrease delays in clinical trial recruitment and minimize human resources that manually review records for prescreening. Weng et al. showed that a real-time screening alert for a single clinical trial before a patient clinical encounter improved the rate of recruitment over a twelve-month period by identifying patients that are “potentially eligible” for full screening. The challenge is to prescreen a patient for an entire portfolio of clinical trials.
2. Related work

Most prior efforts have focused on the standardization of eligibility criteria with expressive ontologies. For example, the Eligibility Rule Grammar Ontology (ERGO) provides a formal representation for eligibility criteria. ERGO is not sensitive to the complexity of criteria and does not represent the stability of eligibility data elements or support fuzzy representations such as “potentially eligible”. A related effort called the eligibility criteria extraction and representation (EliXR) extended ERGO by providing richer semantic information and decomposing complex eligibility criteria into more meaningful semantic segments. EliXR-Time further extended EliXR by supporting temporal representations. For example, a temporal constraint might require a medication washout period of three weeks. Such efforts have focused on rich semantics and the full expressiveness of the criteria, though none of them has been widely adopted yet or have generated real-world impact on clinical trial recruitment processes. In this study, we took a novel approach to this problem by addressing a real user need for classifying major eligibility criteria, which has been observed in folk practice of clinical trial prescreening and has remained largely unstudied.

3. Challenges of cancer clinical trial recruitment

Unique challenges exist for cancer clinical trials. The knowledge of cancer is increasing at an unprecedented rate along with its understood complexity. Patients that can be recruited for cancer clinical trials vary widely in terms of comorbidities, overall health, and cancer pathology. Previous studies have indicated that eligibility criteria concepts may be disease specific and differences in patterns of concepts for eligibility criteria were found between breast cancer and cancer in general. As a result, eligible patients must meet complex criteria, which are not standardized and vary by disease. In addition, not all cancer patients react similarly to the same therapeutic treatment. Heterogeneity in patient response to cancer therapies likely is caused by inter-individual differences in drug deposition and pharmacokinetics. As we enter the era of precision medicine, patients will not be assigned to drug trials without attempting to predict how they will respond to the therapeutic regimen. This complex problem will require efficient information systems that will need to be built within the clinic workflow. Major eligibility criteria offer one effective strategy for streamlining clinical trial recruitment in cancer centers and to potentially increase patient satisfaction by limiting wait-times and maximizing care quality by quickly presenting all available treatment options to patients during clinical encounters.

4. Our previous pilot study

We previously conducted a pilot study involving nine clinical researchers, including physicians, nurses, and clinical research coordinators. Participants were invited to take a survey about how they use eligibility criteria to screen patients for clinical trials. Our study revealed a concept known as “major eligibility criteria”, which could be useful to assist investigators by automatically prescreening patients with data kept in the electronic health record. However, the responses revealed no clear consensus for the definition of “major eligibility criteria”. Each expert panelist defined the concept slightly differently, yet we identified frequently cited major criteria, such as histology and disease stage of cancer, in the aggregate data. Trends also emerged regarding how the panelists used major eligibility criteria in practice. For example, panelists described how they use major eligibility criteria to prescreen patients. The results also uncovered disparities about which criteria should be considered major. Our pilot study helped us to identify the unanswered research question and informed this study, which is to perform in-depth interviews and a focus group to understand the design requirements for prescreening with major eligibility criteria. The pilot study also suggested that major eligibility criteria can potentially improve triage efficiency for clinical trial recruitment by identifying patients that are potentially eligible for clinical trials. However, to the best of our knowledge, no prior study has investigated major eligibility criteria for clinical research recruitment and discussed user requirements for this type of decision support. This study employs a mixed-methods approach that includes both qualitative and quantitative measures to understand how clinical researchers define, select, and use major eligibility criteria. Moreover, we harness their collective knowledge to inform future informatics designs for triaging patients for clinical research eligibility prescreening.

Methods

1. Study setting

We recruited physicians, nurses, and study coordinators from the Breast Cancer Program in the Herbert Irving Comprehensive Cancer Center (HICCC) at Columbia University Medical Center (CUMC). CUMC is an academic medical center in New York City that is affiliated with the New York-Presbyterian Hospital. Clinical researchers in the Breast Cancer Program at HICCC were sent an IRB-approved recruitment email with details about the study and invited to request more information. The protocol activities consisted of one initial interview that was audio recorded, one focus group session, and one survey after the focus group session that study participants completed independently. The invitees were informed that completing the survey was voluntary and they could opt out if they...
did not wish to participate. Eleven people requested information and were subsequently consented to the research study. We recruited eleven clinical research professionals, including five physicians, two nurse practitioners, one nurse, and three research coordinators. Two participants were male and nine were female. Their work experience with breast cancer clinical trials was an average of 6.8 years with a median of 6 years. All participants had experience in recruiting patients for breast cancer clinical trials. Figure 1 graphically depicts our study schema.

![Figure 1. The study schema.](image)

2. The eligibility criteria dataset

Thirty clinical trials were randomly selected from the portfolio of active interventional breast cancer clinical trials, which represented over 41% of the 72 available cancer trials at the point of study at CUMC. Our random sample included a mixture of protocol phases to create a representative sample. All of our selected trials purposely involve interventions done to the patient. Twenty trials were designated as the training dataset, while the remaining ten were the testing dataset. Since significant differences exist in the content and word counts between full text protocols and their summaries on ClinicalTrials.gov, which casts doubt on the reliability of clinical trial summaries, we decided to use eligibility criteria from full-text clinical trial protocol documents for this study.

3. The initial interviews

Ten research members participated in the initial interview to review eligibility criteria for six clinical trials and classified each criterion as either major or minor, so that each trial had three raters. Participants were given the title and all eligibility criteria for each clinical trial that they rated. The response rate was 100%. The survey was conducted using Qualtrics, an online survey platform (www.qualtrics.com). Participants were encouraged to “think aloud” while they completed the task so we could capture their thought process with audio recording, which was later transcribed. All identifying information was stripped from the audio recordings to ensure anonymity. The answers were downloaded from the Qualtrics website into Excel for analysis. The answers for the major eligibility criteria were analyzed individually for themes and then annotated by author MP, who is a subject matter expert in breast cancer clinical trial data collection with over nine years of experience. The initial interview completion rate was 100%. We assigned raters using a cascading structure to maximize variability in rater assignments, which ensured that each rater works with different participants on each clinical trial to maximize result validity.

4. The focus group

The focus group was held one week after completing the initial interviews. Eight participants, including five physicians, a nurse practitioner, a nurse, and a study coordinator, participated in the focus group. The focus group method was used to find areas of agreement and disagreement and participants were encouraged to have a discussion among themselves to complete the tasks. The participation rate of the focus group was 100%. The group started the session by drawing their perception of how the current pre-screening process works. The participants worked individually and were instructed to focus on how physicians, nurses, and study coordinators interact with one another during the prescreening process, if at all. Participants were also asked how, when, and where prescreening occurs.

Next, the group reviewed eligibility criteria that were considered major from the initial interviews using two steps. First, the results of selected trials were reviewed with all focus group participants at the same time. During the discussion, Author MP listed the qualities and attributes of major eligibility criteria on an easel pad to learn why participants considered some eligibility criteria to be major. Author MP updated the list as the conversation progressed during the focus group session. Second, the focus group reviewed their discussion and agreed upon a list of attributes for major eligibility criteria based upon their subject matter expertise and the discussion of the focus group. This data was recorded using an easel pad that was visible throughout the focus group session, where anyone could write a quality or attribute about major eligibility criteria during the conversation. Then, the focus group...
reviewed the attributes and came to a consensus on a final list of attributes. Finally, the focus group participants discussed and drew their version of an optimized prescreening scenario as a group using easel pads that were hung on the wall so all focus group members could participate.

5. The follow-up survey

The follow-up survey was designed to evaluate the standardized major eligibility criteria and desiderata that were derived from our content analysis of the initial interviews and focus group session. Our content analysis was completed in two steps. First, we extracted all eligibility criteria from the training data set of twenty breast cancer clinical trials that were considered major by all three raters and referred to them as consensus major eligibility criteria. Second, we reviewed the list of attributes created by the focus group for major eligibility criteria. To ensure saturation of major eligibility criteria attributes, both authors reviewed and annotated the transcripts of the initial interviews until saturation was reached. Both authors reviewed the interview transcripts separately and discussed their results before coming to a consensus on major eligibility criteria attributes. Differences were resolved by discussion among authors. The result was the desiderata for major eligibility criteria in breast cancer clinical trials.

Ten participants completed the follow-up survey two weeks after the focus group session. The follow-up survey was administered using the Qualtrics survey tool and consisted of three parts. First, study participants were asked to classify each standardized major eligibility criterion that was created from the initial interviews using a 5-point Likert scale with options “Definitely Major”, “Possibly Major-Depending on the Context”, “Undecided”, “Possibly Minor”, and “Definitely Minor”. The second part of the follow-up survey asked study participants to match examples of standardized eligibility criteria with each attribute of our derived major eligibility criteria desiderata. From the training dataset, twelve major eligibility criteria appeared in the training dataset more than once, so we included them in follow-up survey to have study participants match them with the derived desiderata. Study participants were directed to select as many or as few examples of major eligibility criteria for each attribute in our desiderata. The final part of the follow-up survey asked study participants to classify eligibility criteria as major or minor. Ten breast cancer clinical trials were randomly assigned to our ten participants so each participant would classify the eligibility criteria as major or minor for three testing clinical trials. We annotated the major eligibility criteria collected in the follow-up survey to see if any additional standardized major eligibility criteria were detected.

6. Data collection and analysis

Study data was collected with three methods: an online survey tool, Qualtrics, where the results were downloaded into Excel; audio recording and subsequent transcription; and the focus group session that created individual and group prescreening workflow artifacts and easel pad drawings. We utilized the “Think-aloud” Protocol to record the thought process of the study participants while they classified the training set eligibility criteria as either major or minor in the initial interview. Then, we analyzed and annotated the audio transcriptions for the initial interviews using grounded theory to ensure saturation of the desiderata for major eligibility criteria from the focus group easel pads. Finally, we asked study participants to evaluate the standardized major eligibility criteria as well as the desiderata in the follow-up survey.

Results

1. Qualitative study results

The average, median, maximum, and minimum time spent on the interview per participant was 22 minutes 42 seconds, 19 minutes 48 seconds, 40 minutes and 1 second, and 9 minutes and 42 seconds, respectively. Our interview transcript annotation revealed that all study participants differentiated major and minor eligibility criteria. There was 100% agreement amongst the participants that not all eligibility criteria should be given equal weight in the prescreening process. Participant 07 stated, “In my mind major eligibility criteria, I mean like who’s actually worth approaching upfront”. All study participants were able to find one or more major eligibility criteria for every clinical trial that they rated. Several major eligibility criteria were noteworthy and were reported below.

Disease staging was considered a major eligibility criterion for 100% of the study participants. Disease staging appears to be the most important factor to divide patients into subgroups for treatment decisions and recruitment for clinical trials. Participant 03 stated, “So, the disease staging, I would consider major. That’s your number one.” Participant 05 stated, “staging, this is probably one of the most major, most important”. This finding was confirmed by both transcription annotation and the follow-up survey, where 100% of participants identified disease stage as “Definitely Major”.

Transcript annotation also revealed a contextual sentiment about age. Several study participants stated that age could be considered a major or minor eligibility criterion depending on the context, such as how the numerical value of the age creates subgroups in the overall breast cancer population. For example, the age criterion can be major when its
threshold is meaningful, such as “greater than 50 years old”. In contrast, the age criterion “adults greater than 18 years old” is not major because the majority of the patient population satisfies this criterion. Participant 04 stated, “I’m assuming that there’s going to be (over 18)”. Participant 09 stated, “…most of the patients that we see are over 18, so that’s usually an assumption I make that I would call minor, definitely”. The reasoning for this is the degree to which the age value divides the overall breast cancer population into meaningful subgroups. Participant 03 stated, “Most people at least in this setting are going to be 18. When the cut off is larger, like 50, it’s when I would consider it more of a major (criterion)”. 

Physician, nurse, and study coordinator participants did not consider rare phenomena to be major eligibility criteria. Regarding previous cancers, participant 06 stated, “…it wouldn’t be considered a major criterion because for most patients, that’s not a major issue”. Participant 01 speaking about Hepatitis and HIV stated, “While that’s important, it’s not common”. Participant 03 stated, “Bilateral malignancy is relatively uncommon, I would consider that minor”. Participant 09 stated, “I would call that minor because I would say that the majority of the population is not HIV-positive”. Participant 02 stated, “The things that are less common become in my mind not major”. 

An area of contention in the transcript annotation was laboratory results. Specifically, a difference existed between a study coordinator and nurses. Participant 08, a study coordinator, stated, “I can assess the labs…so that’s a major to me”. Participant 09, a nurse practitioner, stated, “I would say in terms of pre-screening, labs are minor because the lab is usually the one we use when they actually come in for their screening visit”. Participant 07, a nurse practitioner, stated, “let’s say the patient is essentially eligible except for some lab variations, for the most part in my experience, this can be remedied”. However, the importance of laboratory results may increase for patients with advanced disease. Participant 07 stated “I feel as though the laboratory parameters…have more implications for patients who are treated in a metastatic setting because those women tend to be more sick”. 

A major eligibility criterion that generated significant conversation in the focus group session was around the concept of “prior treatment”. While prior treatment was not always considered to be accessible in the electronic health record, the focus group did consider it a major eligibility criterion that was available. However, it might take additional time to receive records from other clinicians that treated the patient. Focus group participants clarified that a patient’s prior treatment can be difficult to find, especially if they have been treated at multiple institutions. The focus group stated that the importance of prior treatment differs between early stage and advanced stage breast cancer patients and hence whether prior treatment is major is also a contextual decision. 

The disease staging of the patient may increase the likelihood of other eligibility criteria being considered major. For example, the focus group agreed that the importance of prior treatment proportionally increases in connection with the number of patient prior treatments. For a patient with multiple prior treatments, it is likely that prior treatment becomes a major eligibility criterion for that specific patient. Similarly, if a patient has not had any prior treatments, the concept of prior treatment is less important for that individual patient. This sentiment was also reflected in the transcription annotation. Participant 02 stated during the initial interview, “Given that this is a neoadjuvant study, most patients would not have had prior therapy. So probably not so important to know that they haven’t had prior treatment”. Therefore, prior treatment is more relevant to patients with advanced disease than for early stage disease. 

2. Desiderata for major eligibility criteria 

Desiderata for major eligibility criteria (Table 1) were derived based upon the transcript annotation from the interviews using grounded theory as well as from a group activity that was conducted during the focus group session. The desiderata were then evaluated in the follow-up survey. The desiderata attributes are ranked in descending order of frequency determined by counting the matched eligibility criteria for each term in the desiderata for all participants that completed the follow-up survey. The twelve most frequent major eligibility criteria from the training data set were used for the major eligibility criteria examples. In addition, the major eligibility criterion with the highest score from study participants is identified as a “Best Criterion Example”, where we have included more than one criterion in the event of a tie. 

3. The prescreening process 

During the initial interviews, participants identified different phases of patient screening. Participant 06 stated that major eligibility criteria are used for prescreening, while minor criteria will be used during manual reviews of data to give higher scrutiny to the patient. An example of a minor criterion is asking a patient if they would consider participating in a research trial. The transcript annotation revealed that different phases of patient screening occur at different time points. Participant 06 identified three different steps to screening. The first step is looking at “high level variables”. The second step is when you give a patient a more in-depth look for obvious problems that would prevent them from enrolling in the clinical trial such as organ failure. Thirdly, you have the official eligibility checklist, where you go line by line and see if they meet all eligibility criteria.
At the beginning of the focus group session, study participants were asked to draw the workflow of how patients are currently prescreened in their clinics. 100% of the focus group participants completed the task. Seven out of eight participants identified the first step of the prescreening process starting with the physician, where the physician directs nurses or study coordinators to do a manual screening of the patient. The one remaining participant stated that the prescreening could start with any clinical researcher, depending on the type of study. For example, study coordinators may recruit participants for behavioral or observational studies without physician input.

Table 1. Desiderata of major eligibility criteria for breast cancer clinical trials in descending order of frequency.

<table>
<thead>
<tr>
<th>Desiderata</th>
<th>Operational Definition</th>
<th>Rationale Quotes from Study Participants</th>
<th>Best Criterion Example in the Follow-up Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable</td>
<td>A good biomarker</td>
<td>&quot;I guess the question is how accurate are the data that are going to be in the EMR.&quot;</td>
<td>Disease staging, HER2 status; Hormone receptor status</td>
</tr>
<tr>
<td>Objective</td>
<td>Assessed in a standardized way without opinions</td>
<td>&quot;one way for me to delineate majors versus minors, ...something that’s like blatantly objective, it’s major.&quot;</td>
<td>Age</td>
</tr>
<tr>
<td>Relevant to me</td>
<td>Something I can review on my own</td>
<td>&quot;that’s going to be pathology review, and I don’t have experience on that&quot;</td>
<td>Eligible to receive treatment</td>
</tr>
<tr>
<td>Available</td>
<td>Information ready to use</td>
<td>&quot;there’s always things that aren’t going to be in the medical record or things that you’re not going to know.&quot;</td>
<td>Disease staging</td>
</tr>
<tr>
<td>Stable</td>
<td>No fluctuation in the short term</td>
<td>&quot;If there is something that the patients can discontinue prior to randomization, then that would be minor.&quot;</td>
<td>Diagnosis made by core biopsy; Gender</td>
</tr>
<tr>
<td>Intrinsic patient trait</td>
<td>Something that cannot be changed by external force</td>
<td>&quot;let’s say the patient is essentially eligible except for some lab variations, for the most part in my experience, this can be remedied.&quot;</td>
<td>Age; Gender</td>
</tr>
<tr>
<td>Accessible</td>
<td>Understandable with little effort</td>
<td>&quot;things that would be easy to prescreen&quot;</td>
<td>Age</td>
</tr>
<tr>
<td>Relevant to now</td>
<td>Something I can find out now</td>
<td>&quot;that won’t happen until we actually meet them.&quot;</td>
<td>Age; Gender</td>
</tr>
<tr>
<td>Prevalent in population</td>
<td>Concepts common to breast cancer patients</td>
<td>&quot;wouldn’t be considered a major criterion because for most [breast cancer] patients that’s not a major issue&quot;</td>
<td>Disease staging</td>
</tr>
<tr>
<td>Must have</td>
<td>Not optional</td>
<td>&quot;Must be post-menopausal, major.&quot;</td>
<td>Disease staging</td>
</tr>
<tr>
<td>Differential for population</td>
<td>Able to divide population into subgroups</td>
<td>&quot;...age &gt; 18 is not a major criterion but specific age constraints such as between 50 and 75 are major criteria ”</td>
<td>Disease staging</td>
</tr>
<tr>
<td>Plausible (to happen)</td>
<td>Something realistic and practical</td>
<td>&quot;We would never approach a person with life threatening metastasis&quot;</td>
<td>Eligible to receive treatment; Scheduled to receive treatment</td>
</tr>
</tbody>
</table>

The final group activity of the focus group was to describe an optimized prescreening workflow. The prescreening process utilizes major eligibility criteria to automatically identify potentially eligible patients in the electronic health record. Study participants came to a consensus that decision support for clinical trials must be built into the everyday workflow of the clinicians. At CUMC, the clinicians view a patient schedule that is built for each physician’s clinic day. The focus group agreed that research staff should then be able to easily see potentially eligible patients in the physician’s daily schedule. This enables research staff to screen patients and communicate with one another to make an ultimate screening decision.
Figure 2 shows the optimized prescreening workflow as developed by consensus during the focus group session. In addition to describing the workflow, focus group participants also illustrated the type of information that would be useful in the user interface (Yes; No; Maybe). The relevant clinical trials for each patient are listed on the physician schedule. For example, Patient 1 is scheduled to see the physician at 8:00 AM and is potentially eligible for “Trial A” and “Trial B” as seen in Figure 2. The automated prescreening process, study coordinator, research nurse, and physician can each select “No–Not Eligible, “Yes–Possibly Eligible”, and “Maybe Eligible” after their respective reviews.

4. Quantitative study results

The quantitative results include the instance counts from the training and testing datasets as well as the Likert scale measurements from the follow-up survey. For the training dataset that composed twenty breast cancer trials, the average count of eligibility criteria per trial was 19.7 with a median of 18. There was more than double the amount of criteria that were classified as minor than were classified major. The average major criterion per clinical trial in the training set was 3.55. The average minor criterion per clinical trial in the training data set was 7.3. Similarly, the testing dataset had an average count of 22.5 total eligibility criteria with a median of 20. We again saw there was more than double the amount of criteria that were classified as minor than were classified major. The average major criterion per clinical trial in the training set was 3.2. The average minor criterion per clinical trial in the training data set was 6.9. We annotated the study participant agreed upon major eligibility criteria from the initial interviews as well as from the follow-up surveys. We did not detect any new standardized major eligibility criteria in the follow-up survey. The instance count of the training and testing datasets are seen in Table 2 sorted by total dataset instances. Study participants completed a 5-point Likert scale to classify each major eligibility criterion as “Definitely Major”; “Possibly Major–Depending on the Context”; “Undecided”; “Possibly Minor”; and “Definitely Minor”. We report the mode from participant responses from the follow-up survey Likert scale in Table 2.

Table 2. Standardized major eligibility criteria with instance counts and Likert scale mode.

<table>
<thead>
<tr>
<th>Major Eligibility Criterion</th>
<th>Training Instances</th>
<th>Testing Instances</th>
<th>Total Dataset Instances</th>
<th>Likert Scale Mode for the Participant Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease staging</td>
<td>20</td>
<td>12</td>
<td>32</td>
<td>Definitely Major</td>
</tr>
<tr>
<td>Prior treatment</td>
<td>17</td>
<td>8</td>
<td>25</td>
<td>Definitely Major</td>
</tr>
<tr>
<td>Hormone receptor status</td>
<td>9</td>
<td>3</td>
<td>12</td>
<td>Definitely Major</td>
</tr>
<tr>
<td>Gender</td>
<td>7</td>
<td>2</td>
<td>9</td>
<td>Possibly Major</td>
</tr>
<tr>
<td>Histological confirmed</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td>Definitely Major</td>
</tr>
<tr>
<td>HER2 status</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>Definitely Major</td>
</tr>
<tr>
<td>Pre/Postmenopausal status</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>Possibly Major</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>Possibly Major</td>
</tr>
<tr>
<td>Eligible to receive treatment</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>Definitely Major</td>
</tr>
<tr>
<td>Diagnosis made by core needle biopsy</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>Definitely Major</td>
</tr>
<tr>
<td>Age</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>Possibly Major</td>
</tr>
<tr>
<td>Laboratory values</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Possibly Major</td>
</tr>
<tr>
<td>At least one breast available</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>Possibly Major</td>
</tr>
<tr>
<td>Scheduled to receive treatment</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>Possibly Minor</td>
</tr>
<tr>
<td>Measurable disease</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Definitely Major</td>
</tr>
<tr>
<td>Future scheduled treatment</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Possibly Minor</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Possibly Major</td>
</tr>
<tr>
<td>Language</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Undecided</td>
</tr>
<tr>
<td>Heart rate</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Possibly Major</td>
</tr>
<tr>
<td>Family history of breast cancer</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Possibly Minor</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Possibly Minor</td>
</tr>
<tr>
<td>Suitable to undergo MRI</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Possibly Minor</td>
</tr>
<tr>
<td>Have mobile phone</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Definitely Minor</td>
</tr>
</tbody>
</table>

Figure 3 below displays major eligibility criteria as ranked by study participants in the final survey in decreasing order of significance. The criteria are ranked by significance score, which was calculated by summing all participant scores for each criterion. The criteria received the following scores for each classification: “Definitely Major”=1; “Possibly Major–Depending on the Context”=2; “Undecided”=3; “Possibly Minor”=4; and “Definitely Minor”. The criteria are ranked by significance score, which was calculated by summing all participant scores for each criterion. The criteria received the following scores for each classification: “Definitely Major”=1; “Possibly Major–Depending on the Context”=2; “Undecided”=3; “Possibly Minor”=4; and “Definitely Minor”.

2031
Minor”=5. The criterion with the lowest score is considered the most significant. For example, “Disease staging” had the lowest significance score because all participants identified it as “Definitely Major”.

**Figure 3.** Major eligibility criteria measured with a Likert scale by participants in the follow-up surveys.

**Discussion**

Our training dataset was considered a representative sample of interventional breast cancer clinical trials at CUMC for three reasons. First, our results showed that “Hormone receptor status” was identified as the third most common major eligibility criterion with 12 total instances in our dataset and second most significant major eligibility criterion after disease staging. This observation confirmed sample representativeness since the prevalence of hormone receptor positive breast cancer in the entire breast cancer population is approximately two-thirds of all reported cases. Second, saturation was reached in our training dataset of major eligibility criteria instances since no new major eligibility criteria were detected in our testing dataset. Third, the sample represented over 41% of interventional breast cancer trials available at CUMC. The results of our mixed-methods study prove the existence of major eligibility criteria. In addition, both the training and testing datasets had similar summary statistics. The average count of total eligibility criteria for both training and testing datasets was 19.7 and 22.5, respectively. The average count of major eligibility criteria for training and testing data sets was 3.55 and 3.2, respectively. Minor eligibility criteria are more prevalent in our sample of clinical trials than major criteria. This finding is significant because automating a system that needs to search for fewer machine-readable standardized criteria is more efficient and feasible operationally. In addition, this finding questions prior efforts to provide a complex knowledge representation of highly expressive concepts that gives equal weight to all eligibility criteria. Our study suggests that time could be saved by focusing on major criteria that were identified by study participants. It is possible that raters identify minor criteria more easily than major criteria, but an inescapable conclusion is that complex knowledge representations that give equal weight to all criteria are perhaps less useful than previously thought. This is an area that requires further exploration.

We have derived a standardized list of major eligibility criteria for breast cancer clinical trials at CUMC. We have confidence in our list being complete because we did not detect any new standardized major eligibility criteria in our testing dataset during the follow-up survey. While we focused exclusively on breast cancer, our study provides a mixed-methods approach for discovering major eligibility concepts in other diseases. This approach will be helpful in future projects since clinical trials lack standardized eligibility criteria. As a consequence, recognizing similar criteria across two or more clinical trials remains a challenge even in the same disease realm.

Our approach differs from previous research attempts to create eligibility criteria knowledge representations by avoiding highly expressive standardization and temporal restraints in favor of focusing on only the most user-relevant concepts. The granularity of the major eligibility criteria desiderata is coarser than other knowledge representations and more comparable to available electronic health record data. The coarseness of our standardized criteria is a very strong advantage for the usefulness of major eligibility criteria because it takes into account practical considerations such as data availability in electronic health records and fitness to the clinic workflow. Our
focus group detailed how clinical research professionals use major eligibility criteria in their thought process while prescreening patients. Focus group data provide insight into building a clinical decision support system that is powered by major eligibility criteria to automatically suggest clinical trials that patients are potentially eligible for during clinical encounters.

We also sought to understand the rationale of major eligibility criteria. The study participants viewed HIV as less important for breast cancer clinical trial recruitment because it is rare among the target population. Other criteria, such as age, depend on context. Annotated transcripts from the initial interviews, the focus group session, and the follow-up survey provide evidence that prior treatment is significant, yet difficult to obtain. Thus, prior treatment is difficult to utilize in automatic prescreening. More advanced disease staging may increase the importance of some eligibility criteria such as laboratory results and prior treatment. As a consequence, clinical trials for early stage breast cancer could realize more of a benefit using major eligibility criteria for automated prescreening than studies for advanced disease stages.

The results of our mixed-methods study also identify the desiderata of major eligibility criteria for breast cancer clinical trials. While we utilized only breast cancer clinical trials, our desiderata could be scalable to other diseases. Furthermore, the desiderata could provide guidance about what features will create an optimal clinical trials recruitment tool for future clinical decision support systems. Since the desiderata terms are ranked by frequency, they provide weighted insight into how clinical researchers think while prescreening patients for clinical trials. For example, “Reliable” was the desiderata attribute that was most frequent to our major eligibility criteria followed by “Objective” and “Relevant to me”. The relatively highly ranked “Relevant to me” desiderata attribute suggests that a role-based analysis may provide additional insight into prescreening processes.

Limitations
Since the results are only from one academic medical center, we cannot speculate to the universality of the desiderata. We purposely included only breast cancer clinical trials in our dataset because previous research findings suggest that relevant eligibility criteria are disease specific. The applicability of our findings to other diseases is unknown. This study did not use any real patient data, instead relying on observing the opinions and thought processes of clinical research professionals using the “Think-aloud” protocol. As a consequence, we are unsure what effect a fragmented or incomplete patient electronic health record would have on influencing the major eligibility criteria desiderata. Finally, even though our study methods included qualitative and quantitative methods, our small sample size prohibits us from making definitive conclusions about breast cancer clinical trials in a larger context.

Future work
The scalability of the major eligibility criteria desiderata to other diseases and institutions is an unanswered research question that warrants additional exploration. Our desiderata may assist clinical researchers and informaticists to make informed decisions about prioritizing the standardization and collection of data in the electronic health record for clinical trial prescreening. 100% of study participants agreed disease stage was “Definitely Major” during the follow-up survey. This suggests that we should prioritize the capture of this information in a standardized, machine-readable way for efficient prescreening. Similarly, “Hormone receptor status” and “HER2 status” should be prioritized for standardized capture. Standardized major eligibility criteria could provide a basis for clinical decision support by triaging patients for clinical trial prescreening and other applications in personalized medicine.

Future work should evaluate how major eligibility criteria are captured in electronic health records and how prescreening results can be best presented to clinical researchers in the clinic workflow. Our study has shown how major eligibility criteria might be used in an optimized workflow to screen patients for clinical trials. Automated prescreening could reduce human effort and expense by reducing the manual review of patient charts. If financial or other barriers prevent major eligibility criteria from being collected in a standardized way or even preclude an automated prescreening process, the benefits of major eligibility criteria still exist. For example, major eligibility criteria can still inform the development of manual prescreening methods. Additional evaluation of the desiderata and standardized major eligibility criteria will be helpful for understanding how our research can be utilized when prescreening patients in real life scenarios.

Conclusion
Our results demonstrate that major eligibility criteria exist for breast cancer clinical trials. This study investigated how, when, and why clinical researchers use major eligibility criteria to triage patients for clinical trial prescreening. Our identified desiderata can potentially inform future design of clinical decision support for clinical trial prescreening. Future research is warranted to test the generalizability of our method beyond breast cancer trials and consider real patient data in order to gain insight into how the major eligibility criteria desiderata is influenced by fragmented or incomplete patient records and what steps might be taken to mitigate those effects.
Acknowledgments
This study was sponsored by the U.S. National Library of Medicine grant R01LM009886 (PI: Weng) and U.S. National Center for Advancing Translational Science grant UL1 TR000040 (PI: Ginsberg).

References
An Assessment of Family History Information Captured in an Electronic Health Record

Fernanda Polubriaginof, MD1, Nicholas P. Tatonetti, PhD,1 David K. Vawdrey, PhD2,1
1Department of Biomedical Informatics, Columbia University, New York, NY, USA;
2NewYork-Presbyterian Hospital, New York, NY, USA

Abstract

Family history is considered a core element of clinical care. In this study we assessed the quality of family history data captured in an established commercial electronic health record (EHR) at a large academic medical center. Because the EHR had no centralized location to store family history information, it was collected as part of clinical notes in structured or free-text format. We analyzed differences between 10,000 free-text and 9,121 structured family history observations. Each observation was classified according to disease presence/absence and family member affected (e.g., father, mother, etc.). The structured notes did not collect a complete family history as defined by standards endorsed by the U.S. Agency for Healthcare Research and Quality: the free-text notes contained more information than the structured notes, but still not enough to be considered “complete.” Several barriers remain for collecting complete, useful family history data in electronic health records.

Introduction

Family history has always been considered “a core element of clinical care.”1 It plays an important role in personalized medicine, being a free genetic tool that almost every patient has access to.2 Since the Human Genome Project, new genomic tools have been described,3 but family history is critical for identifying patients that may benefit from more intensive screening. Family history provides information that enables individualized disease diagnosis, treatment and prevention.

Several studies have shown that family history can help predict the risk of a patient developing certain diseases such as breast cancer, colorectal cancer, ovarian cancer, osteoporosis, cardiovascular disease, psychiatric disorders, and diabetes. Knowing that a patient is at increased risk of developing a disease based on family history enables disease prevention that can vary from intensive screening to prophylactic surgery, early diagnosis and/or early and tailored treatment. For example, current guidelines from the American Cancer Society define criteria for MRI eligibility in addition to mammography for breast cancer screening4,5. The guidelines recommend that patients that have lifetime risk of breast cancer greater than 20% by BRCAPRO6, Tyrer-Cuzick7 or Claus8 models should have screening MRI in addition to mammography. Each of these models was developed using different methods, different populations and different risk factors and each of them was developed to predict different outcomes, but all of them heavily rely on family history and presence of risk factors9. This is just one of many examples of the importance of family history information for clinical care.

Since several guidelines and models exist to estimate a patient’s risk for different conditions, clinical decision support systems have been developed to facilitate guided personalized medicine. One of the earliest research studies about the use of clinical decision support (CDS) systems in personalized medicine used was conducted by Emery and colleagues in 1999. The study identified that the CDS systems available were not appropriate for use in a primary care setting.10 To address this problem, they developed a system to record and interpret family history data in the primary care clinic. The system included family history relevant to breast, ovarian and colorectal cancer.11 Over time, other clinical decision support systems were developed to manage other types of cancers such as colorectal cancer and Lynch syndrome. All these systems used family history information to provide risk assessment.12

Currently, the U.S. Preventive Services Task Force (USPSTF) includes risk assessment based on family history for some conditions, demonstrating the importance of family history in clinical care. Conditions that include family history information as part of the USPSTF are screening for BRCA mutation and BRCA-related cancers13, osteoporosis14 and lipid disorders in adults15,16. A 2007 report commissioned by the Agency for Healthcare Research and Quality (AHRQ) recommended that collection of family history information should include diseases in first-degree relatives and second-degree relatives from both the maternal and paternal side, the relative’s age at the time of disease diagnosis, and each relative’s race and ethnicity.17
In electronic health records (EHRs), family history information may be collected in a structured or narrative (i.e., “free-text”) format. On one hand, structured documentation is ideal for data reuse since it may be coded and standardized. On the other hand, when clinicians are providing care, flexibility, efficiency, quality, expressivity are important. The purpose of this study was to understand how family history data were captured in an established commercial EHR system at a large academic medical center, and to assess the quality of the data collected.

Methods

With Institutional Review Board approval, we conducted a retrospective analysis of data from the Allscripts EHR (Allscripts Corp., Chicago IL) used at NewYork-Presbyterian Hospital/Columbia University Medical Center since 2004. Our study focused on the difference between family history data collected using structured fields vs. free-text. Each note template in the EHR contained one or more “observation” data elements. An observation could be a text box, a Boolean (e.g., a checkbox or radio button), or numeric value. Text boxes could be fully free-text, or they could be constrained to allow only items from a predefined list, such as the words “low,” “medium,” or “high.” Observations had an internal code and description specified using a configuration tool in the EHR. While our EHR vendor provided some predefined observations, the vast majority were locally defined and did not comport with any existing standard terminology.

The EHR system contained 1,560 active templates for documentation. Each of these templates contained one to several hundred discrete observations. There were 140,038 observations defined in the EHR; of those, 653 had an internal code containing the words “fam hx” OR “family hist.” We identified the note templates that contained these observations and queried the EHR database to identify the number of times each note template was used, as well as the number of unique patients who had at least one of these observations recorded. The number of times each note template was used varied from 1 to 79,505, and number of unique patients varied from 1 to 67,276 for each note template. The note templates that contained the most commonly used free-text and structured text observations were selected for analysis.

The most-used note template that contained structured family history observations was the Ambulatory OB/GYN Antepartum Record (Figure 1). This note template was used in our institution for obstetric patients in the institution’s ambulatory care network. Overall, this note template was used 79,505 times for 67,276 unique patients. The most-used free-text family history observation was the Neurology Admission Note (Figure 2). This note was used for every patient admitted to the neurology service. The Neurology Admission Note was used 49,656 times for 22,642 unique patients.

Figure 1. Ambulatory OB/GYN Antepartum Record: the most-used template note that contained structured family history observations.
For both the Ambulatory OB/GYN Antepartum Record and the Neurology Admission Note templates, 10,000 family history observation entries were randomly selected from notes between 2007 and 2014. Manual annotation by a clinical expert (FP) was performed in all 10,000 free-text observations, as well as in all structured observations that occurred more than once (9,121 observations). Categories were identified based on the content of information in the observations and the standards endorsed by AHRQ and were used to annotate both datasets. In total, 19,121 observations were manually annotated and assigned to a category. The categories used were: 1) presence of disease in specified relative(s), 2) presence of disease in unspecified relative(s), 3) absence of disease and 4) other (Table 1). The annotation results were compared between the datasets.

Table 1. Family history categories with definitions and examples.

<table>
<thead>
<tr>
<th>Family History Categories</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of disease in specified relative(s)</td>
<td>When family history of disease is reported paired with a relative</td>
<td>“Mother: hepatic cancer; Brother: colon cancer”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Hypertension Mother”</td>
</tr>
<tr>
<td>Presence of disease in unspecified relative(s)</td>
<td>When a family history of disease is reported by itself without affected</td>
<td>“History of diabetes, hypertension, MI, strokes.”</td>
</tr>
<tr>
<td></td>
<td>relative information</td>
<td>“Diabetes”</td>
</tr>
<tr>
<td>Absence of disease</td>
<td>When family history of disease is negated</td>
<td>“No dementia; No strokes.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Diabetes Denies”</td>
</tr>
<tr>
<td>Other</td>
<td>Miscellaneous responses</td>
<td>“Non-contributory”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“None”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“no Arabic translator”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“No family at bedside and pt nonverbal”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Pt adopted, unknown family history”</td>
</tr>
</tbody>
</table>
Results

Figure 3 shows the results of the manual annotation of family history observations. In the category “Presence of disease in specified relative(s),” 58.7% of the observations from the Neurology Admission Note (free-text) contained family history information including the disease as well as the relative affected. However, when analyzing data from the observations from the Ambulatory OB/GYN Antepartum Record (structured), only 5.2% contained information specifying the patient’s relative. In contrast, 7.3% of the observations from the Neurology Admission Note (free-text) contained information categorized as “Presence of disease in unspecified relative(s),” and 50.1% of the observations from the Ambulatory OB/GYN Antepartum Record (structured) captured this type of information (Figure 3).

Furthermore, 27.5% of the observations from the Neurology Admission Note (free-text) captured information about the absence of family history of a certain disease, while only 0.9% of the observations from the Ambulatory OB/GYN Antepartum Record (structured) captured information with this level of detail (Figure 3).

A large proportion (39.2%) of the observations from the Ambulatory OB/GYN Antepartum Record (structured) were classified as “Other.” The vast majority of these cases referred to family history described as “N/A.” Such description provides no information of the patient’s family history (Figure 3).

Discussion

When analyzing family history data collected using structured vs. free-text data elements, the clinical expert annotations revealed that there was a considerable difference between the content of the family history information collected. The structured template was shown to be a poor way to collect family history since the relative affected by the condition was not captured. Other important details such as age of onset and vital status were also not being captured. In the template shown in Figure 1, text could be optionally entered in the free-text box but this field was stored separately in the database from the structured component. In Figure 1, it was also unclear what was meant by “N/A.” Possible interpretations were that this could indicate that a patient would not or could not inform, had no knowledge or even that such questions were not asked. Another problem was that rows could easily be left empty and it is unclear what that means. It could mean that the clinician did not ask about the family history of a certain condition, or that the patient did not know. Furthermore, the free-text box was used to capture relatives, deny presence conditions, or even to record other types of information such as age, type of cancer, etc. Notes that used the...
free-text template were more comprehensive and often contained much useful information. However, not all of them contained pertinent and useful family history information.

Unlike some EHR implementations where family history information is kept in a centralized fashion, our institution did not have a centralized location to store this type of data. Family history data was collected and stored as part of the clinical notes. The fact that different healthcare providers collected this type of information in several different notes may have resulted in inconsistencies across notes. Our study discovered that our institution’s EHR 653 observations related to family history. Many different note-writing templates were available and many included family history sections, but there was no standardized method for collecting this type of data. One benefit of having so many different ways of collecting family history data was that we could assess the quality of data based on characteristics such as structured items vs. free-text fields.

We also observed that free-text boxes in both notes were often used to capture information that was unrelated to family history. For example, indicators were included such as: “intubated,” and “no family at bedside and pt nonverbal”. These are important pieces of information about the patient but should not be reported as part of the family history section.

Despite the well-known and well-described importance of family history, several barriers exist in its collection and analysis, as well as in its use for personalized management based on patients’ risk assessment. Barriers to collect family history can be classified in two major categories: clinician-related and patient-related.

Clinician-related barriers include lack of time to obtain, organize and analyze family history information; lack of resources and lack of reimbursement for such activity; underestimation of the value of family history data by the clinician; lack of expertise in obtaining and analyzing family history information; lack of standards for family history collection; and lack of clear guidelines to assess patient risk based on family history. The first, and perhaps the most critical barrier for family history collection is lack of time to obtain, organize and analyze family history information. Obtaining complete and accurate family history information, organizing it in a pedigree and analyzing family history data is extremely time-consuming. Furthermore, it is not sufficient to collect family history from patients only once. It is important to regularly update family history information, analyze it, and reconcile conflicting information. A 1989 study surveying four genetic clinics reported that the time patients spent in the first consultation varied from 3–5.5 hours, with over half of this time spent before or after the patient’s appointment. A 2011 study demonstrated that while the majority of clinicians (77.5%) reported collecting cancer family history on their patients, only 26.0% included minimum adequate cancer family history. Furthermore, 57.4% of clinicians updated family history information just once a year, and 22.2% of clinicians never updated family history information for their patients at all. When questioned about the barriers to collecting cancer family history, clinicians reported lack of time as the primary issue. The study focused on cancer family history, but it demonstrated how challenging family history is to collect and maintain, in general.

Lack of resources and reimbursement for family history collection is another important barrier. Clinicians are not reimbursed for the time spent on family history collection and risk assessment. In fact, in 2009, lack of incentives from the government was being described as one of the challenges prohibiting adequate collection of family history. In addition to misaligned incentives, lack of standards has also been reported as a challenge in this area. The lack of standards for data elements, terminology, structure, interoperability, and clinical decision support rules for family history data is a huge obstacle to implement it in the clinical workflow. This point is underscored by the existence of multiple EHR templates available to assist physicians in capturing family history data. Furthermore, limited knowledge and lack of expertise in obtaining and analyzing family history by clinicians is another barrier that has been described in several studies.

There are also barriers to collecting accurate family history data on the patient side. These include uncertainty about biological family composition; uncertainty about the health history of family members; inaccuracies in patient recall, language-related and cultural factors. Clinicians often cite uncertainty about biological family composition as a challenge when collecting family histories, especially in cases where the patient is part of a large biological family. Language-related and cultural factors can also be a factor that negatively affects collection.

There are now several initiatives to facilitate and encourage the use of family history data. These initiatives are focused on the use of these data for precision medicine, where the need for accurate and detailed family history data is great. Three such initiatives are: Stage 2 of the federal “Meaningful Use” EHR financial incentive program, the U.S. Surgeon General’s “My Family Health Portrait,” and the HL7 Clinical Genomics Family History/Pedigree Model.
Stage 2 of the Meaningful Use program included a requirement of clinician’s to use structured data entry for family history. Eligible hospitals had to have for 20% of their patients at least one structured family history data element, for at least one first-degree relative in the electronic health record. As discussed above, lack of incentives to collect family history is an important barrier. The Meaningful Use program will be a strong incentive for U.S. hospitals to collect family history information. Although the determined measure of at least one structured data entry for at least one first-degree relative is far from what is considered complete family history, it is a start.

Secondly, since family history data collection is extremely time-consuming, innovative tools have been created to facilitate this process. Some are leveraging patient input of family history data. The U.S. Surgeon General’s My Family Health Portrait, a federal initiative to collect family history, is a website that allows patients to collect family history information and share their information with family members and healthcare providers. A study conducted in 2011 described that the average time taken to input family history information was 15 minutes, in a range from 3 to 45 minutes. Instead of having a healthcare provider questioning patients about their family history, patients can enter their own data, saving clinician time—the major barrier for family history data collection. This practice also engage patients in their care and gives them time to review their family information and contact relatives and question them about information that they may not know. Engaging patients in this fashion encourages more accurate family history information. Of note, one advantage of using electronic questionnaires is that certain questions can be made mandatory, and branching logic can be employed. In contrast, in a clinical encounter, the doctor may forget to ask certain questions or may skip questions due to lack of time.

It is important to emphasize that to fully represent family history information, data representation must be multidimensional since it is necessary to not only capture the disease but also the relative affected, age of onset, and cancer type if applicable. Moreover, development of standards to support interoperability is essential for sharing data for clinical care and clinical research purposes. In the domain of family history, HL7 has a workgroup that specifically works on development of models for representing family history. The workgroup has developed the HL7 Clinical Genomics Family History (Pedigree) Model. It is a standard for capturing data within a system as well as to transmit family history data between systems. It includes patient’s family and familial relationships, diseases, genetic data and risk analysis. This HL7 standard is already used by the U.S. Surgeon General’s My Family Health Portrait application, and we believe it will be important for EHR vendors and other stakeholders to adopt this standard.

Limitations
The study analyzed only observations contained in two note templates (out of a total of 1,560 available) in our EHR. One note template was used in the ambulatory care setting and the other template was used for hospital admissions. Although the templates were selected based on the fact that they were the most frequently used templates at our institution, it is unclear if analysis of other EHR templates would yield different conclusions.

Conclusion
In summary, family history information is a valuable tool for personalized medicine. Computer-based tools can improve data collection, risk assessment and decision-making, facilitating effective use of family history in the clinic, engaging patients in their care, and therefore providing better personalized care. Our electronic health record vendors have not developed a user-friendly system to collect this type of data. Patient-facing tools for collecting family history data are not yet being used on a large scale. Numerous efforts have been made to collect family history data in the electronic format and to facilitate its use in the clinical setting, but several barriers remain unsolved. We have demonstrated that free-text observations were more comprehensive than structured observations; however neither was ideal for capturing patients’ complete family history.

References
Computational Methods for Unraveling Temporal Brain Connectivity Data

Bisakha Ray¹, MSE, Alexander Statnikov¹, PhD, Constantin Aliferis¹, MD, PhD
¹Center for Health Informatics and Bioinformatics,
New York University School of Medicine, New York, NY, USA

Abstract

Brain science is a frontier research area with great promise for understanding, preventing, and treating multiple diseases affecting millions of patients. Its key task of reconstructing neuronal brain connectivity poses unique Big Data Analysis challenges distinct from those in clinical or “-omics” domains. Our goal is to understand the strengths and limitations of reconstruction algorithms, measure performance and its determinants, and ultimately enhance performance and applicability. We devised a set of experiments in a well-controlled setting using an established gold-standard based on calcium fluorescence time series recordings of thousands of neurons sampled from a previously validated neuronal model of complex time-varying causal neuronal connections. Following empirical testing of several state-of-the-art reconstruction algorithms, and using the best-performing algorithms, we constructed features of a classifier and predicted the presence or absence of connections using meta-learning. This approach combines information-theoretic, feature construction, and pattern recognition meta-learning methods to considerably improve the Area under ROC curve performance. Our data are very promising toward the feasibility of reliably reconstructing complex neuronal connectivity.

Introduction

Brain science is a frontier research area with great promise for understanding, preventing, and treating multiple diseases affecting millions of patients. With the launching of international initiatives like the BRAIN Project [1], and the Human Brain Initiative [2], comes the daunting task of dismantling the complicated, intertwined circuitry of the human brain. Accurately reconstructing brain structure and mapping it to function remains a formidable analytic challenge, vastly exceeding, at the present time, the difficulty of deciphering the structure and function of metabolic, transcriptomic or proteomic networks. Furthermore, mapping from genes to neurons to neuronal circuits to actual human behavior and intelligence poses a Big Data Science problem of unprecedented difficulty. The research community has been rapidly developing and testing methods addressing the above tasks. For example, recordings of about 80% of the hundreds of thousands of neurons of the larval zebrafish are publicly available awaiting efficient computational techniques to crack the circuit and recover the underlying connectivity network [3]. Reverse engineering of networks, like in the DREAM challenge [4], is common in bioinformatics and is now gaining momentum in neuroscience as well.

A significant pool of methods to draw powerful techniques from is causal discovery methods. These were developed to address the financial and ethical concerns associated with randomized controlled trials. A reflection of their significance is that they have been recognized with both the Turing Award (to Judea Pearl) and the Nobel Prize (to Clive Granger). For example, several variations of Granger causality (GC) have been used for causal structure discovery in neuroscience using different modalities like fMRI data [5] and electrophysiological (EEG) signals [6].

On the experimental validation front, calcium fluorescence imaging signals of neuronal cultures provide temporal, observational data for causal network reconstruction of neurons in vivo. It allows for the simultaneous recordings of thousands of neurons and can be combined with other state of the art technology like optogenetics [7]. This experimental model is less time consuming and less tedious than conventional experimental techniques like axonal tracing [8].

Problem and Approach

As shown in Figure 1, the data for reconstruction consists of calcium fluorescence imaging recordings of the activity of neurons in a given brain network. The aim is to predict the directed connections of the neuronal network from raw data. As experimental data lacks ground truth value to verify the inferred network, we use simulated data from a broadly accepted ground truth model with realistic behavior coming from the Connectomics challenge [9] that aimed to predict the directed connections (synapses) of a neural network, given calcium fluorescence imaging recordings of the activity of each neuron in this network. In a network of \( n \) neurons, there are \( n^2 \) potential directed connections.
Among the $n^2$ possible combinations, for each pair of neurons $(i,j)$, the problem is to determine whether there is a connection $i \rightarrow j$. 180,00 samples (1 hour at 50 frames per second) were generated.

![Image](real Calcium Fluoresce Imaging: Connections?)

Figure 1. Research design of the present study [10].

Causal-Structure Learning Techniques

Several reconstruction methods are used for causal structure learning. The reconstruction strategies can be classified as correlation-based, entropy-based, and GC-based.

Correlation-Based Measures

1. Pearson’s Correlation

Pearson’s correlation coefficient is a measure of the degree of linear association between variables. For variables $X$ and $Y$ with means $\bar{X}$ and $\bar{Y}$, Pearson’s correlation, for $X$ and $Y$ is calculated as in (1),

$$\text{Pearson’s correlation}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

where $n$ is the number of dimensions of $X$ and $Y$. 

2044
2. Cross-Correlation

Cross-correlation involves correlating variables over many different time lags. The shifting correlation in Cross-correlation can be achieved mathematically by introducing a time lag. The Cross-correlation, at time delay, $d$ between two time series $X$ and $Y$, with means $\bar{X}$ and $\bar{Y}$ can be calculated as in (2)

$$\text{Cross-correlation}(X,Y) = \frac{\sum(X_i - \bar{X})(Y_{i-d} - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_{i-d} - \bar{Y})^2}}.$$  

(2)

The above can be computed for all time delays $d = 0, 1, 2, \ldots$ At time delay 0, this reduces to calculating correlation coefficient between time series $X$ and $Y$. Cross-correlation assigns a score to each possible link between two nodes based on the highest cross-correlogram peak for assigned time lags or delays.

Entropy-Based Measures

Entropy is defined as the amount of information change in a random variable or the average uncertainty of a random variable. The entropy $H$ of a discrete random variable $X$ is expressed as in (3)

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x),$$  

(3)

where $p(x)$ is the probability that $X = x$.

The joint entropy of two variables $x$ and $y$ is defined in (4) as

$$H(X,Y) = -\sum_{x \in X} p(x,y) \sum_{y \in Y} p(x,y) \log_2 p(X,Y).$$  

(4)

1. Gini Index

Gini Index is a measure of statistical dispersion or inequality in a population. It measures the divergences between the values of the probability distributions of the target ($Y$) and attribute ($X$).

2. Mutual Information

Mutual Information (MI) expresses the amount of information that one variable contains about another variable. Mutual information is defined as the reduction in uncertainty of one random variable $X$ given the knowledge about another variable $Y$ and is calculated as in (5).

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)},$$  

(5)

where, $p(x,y)$ is the joint probability distribution function of variables $x$ and $y$, $p(x)$ is the marginal distribution functions of $x$, and $p(y)$ is the marginal distribution function of $y$.

This quantity measures the distance between the joint probability density or distributions between variables $x$ and $y$. The mutual information is the difference in the entropy of signal $X$ by itself and the entropy of $X$ when it is conditional on another signal $Y$ or, it is the measure of difference between the joint probability distributions of the variables and their independent distributions.

3. Generalized Transfer Entropy

In its original formulation [11], for two Markov processes $X$ and $Y$ of order $k$, where $x_n^{(k)}$ is a vector of length $k$ whose entries are samples of $X$ at time $n$, $n-1$, $\ldots$, the transfer entropy (TE) from $Y$ to $X$ is defined as in (6),

$$TE_{Y \rightarrow X} = \sum P(x_{n+1}, x_n^{(k)}, y_n^{(k)}) \log \frac{P(x_{n+1}|x_n^{(k)}, y_n^{(k)})}{P(x_{n+1}|x_n^{(k)})},$$  

(6)

From the original paper by Stetter, et al. [12], TE can be defined as the distance between the probability distributions or the Kulback–Leibler distance between the single-node transition matrix $P(x_{n+1}|x_n^{(k)})$ and the two-node transition matrix $P(x_{n+1}, x_n^{(k)}, y_n^{(k)})$. TE is zero if the two probability distributions are identical or the distance between them.
is zero. If it greater than zero, it indicates dependence of \( x \) on past values of \( y \). Synaptic time constants of the neuronal network (~1 ms) are much shorter than the acquisition times of the recording (~10 ms).

In [12], in the formulation of Generalized Transfer Entropy (GTE), causal interactions between same bin nodes or events that occur in the same time slice are accounted for. Also, the network switches between different dynamical states of bursting and non-bursting. The fluorescence histograms of networks 1 and 2 respectively in Figure 2 are right skewed and resemble Gaussians with heavy tails on the positive side. If one were to fit the Gaussian profile (with a Gaussian function) and the tail with another suitable function, the point where they intersect would be the conditioning level. To restrict the evaluation of the network to a consistent range reflective of the overall collective network dynamics, the signal is averaged for the whole network as in (7) [12].

\[
g_t = \frac{1}{N} \sum_{i=1}^{N} x_i(t)
\]  

(7)

All data points in which the average fluorescence \( g_t \) is below this predefined threshold called a conditioning level or \( g \) are included in the analysis. Therefore, the authors restrict themselves to only time points \( t \) that fulfill the condition \( \{ t : g_t < \bar{g} \} \). Hence, TE using the above two considerations is reformulated as in (8) [12].

\[
TE_{Y \rightarrow X}^* (\bar{g}) = \sum P(x_{n+1}^{(k)}, y_{n+1}^{(k)} | y_{n+1}^{(k)} < \bar{g}) \log \frac{P(x_{n+1}^{(k)}, y_{n+1}^{(k)}}{P(x_{n+1}^{(k)})}
\]  

(8)

Granger Causality-Based Method

\( Y \) Granger-causes \( X \) if for some \( s > 0 \), the mean squared error of a forecast of \( X_{t+s} \) based on \( (X_t, X_{t-1}, \ldots) \) is larger than based on \( (Y_t, Y_{t-1}, \ldots) \) and \( (X_t, X_{t-1}, \ldots) \).

\( Y \) fails to Granger-cause \( X \) if for all \( s > 0 \), the mean squared error of a forecast of \( X_{t+s} \) based on \( (X_t, X_{t-1}, \ldots) \) is the same as that is based on \( (Y_t, Y_{t-1}, \ldots) \) and \( (X_t, X_{t-1}, \ldots) \).

Experimental Design

The study consisted of the following steps:

1. Datasets and data preparation

The datasets generated for the experiments used a realistic neuronal connectivity network simulator [8, 12]. The goal of the inference algorithms is to predict whether there is a connection between neuron \( i \) and neuron \( j \) in a network of \( N \) neurons. The resulting connectivity scores matrix is a \( N \times N \) matrix where \( N \) is the number of neurons and each entry \((i,j)\) gives an estimate of the strength of the connection, \( i \rightarrow j \). The original ground truth network structure is provided in another similar \( N \times N \) matrix, where entry \((i,j)\) indicates connection between neurons \( i \) and \( j \). Training and test data for networks of around 100–500 neurons are generated using the neuronal connectivity simulator. Table 1 provides a description of the generated networks. The connection and inhibitory probability are based on similar studies from literature [12].

<table>
<thead>
<tr>
<th>Network Identifier</th>
<th>Number of nodes</th>
<th>Connection Probability</th>
<th>Inhibitory Probability</th>
<th>Clustering Coefficient</th>
<th>Indegree Median [95% Interval]</th>
<th>Outdegree Median [95% Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network 1</td>
<td>100</td>
<td>0.12</td>
<td>0.2</td>
<td>0.6</td>
<td>12 [7, 18]</td>
<td>11 [5, 20]</td>
</tr>
<tr>
<td>Network 2</td>
<td>100</td>
<td>0.16</td>
<td>0.2</td>
<td>0.6</td>
<td>16 [10, 25]</td>
<td>16 [9, 24]</td>
</tr>
<tr>
<td>Network 3</td>
<td>500</td>
<td>0.12</td>
<td>0.2</td>
<td>0.16</td>
<td>59 [46, 74]</td>
<td>59 [47, 74]</td>
</tr>
<tr>
<td>Network 4</td>
<td>500</td>
<td>0.16</td>
<td>0.2</td>
<td>0.18</td>
<td>79 [64, 97]</td>
<td>79 [65, 96]</td>
</tr>
</tbody>
</table>
Using the NEST [13] simulator, realistic dynamics like light scattering artifacts, bursting activity, and spontaneous firing dynamics of real, excitatory integrate-and-fire neurons are incorporated in the networks. In order to analyze the performance of the algorithms and benchmark them, we generated synthetic calcium fluorescence time series from the simulated cultured neural network that exhibits realistic dynamics. Since the surrogate ground truth topology of cultures is also available, the quality of the reconstruction can be evaluated by comparing the inferred with the real network connectivity.

A random pair of neurons is selected and their connections are crossed while maintaining the indegree and outdegree of the nodes. The clustering coefficient (CC) of a network by Watts and Strogatz [14] is defined as follows. Let us suppose a vertex V in a graph has \( k \) neighbors. If every vertex is connected to every other vertex in the graph then a maximum of \( \binom{k}{2} = \frac{k(k-1)}{2} \) edges can exist between them. Let \( C^' \) be the fraction of these possible edges that actually exist between a pair of nodes. The CC is then the average of \( C^' \) over all V. The crossing process of the connections was repeated until a desired CC was obtained. The diagonal entries of the connectivity matrix representing a neuron’s connections to itself are removed.

The clustering coefficient (CC) of a network by Watts and Strogatz [14] is defined as follows. Let us suppose a vertex V in a graph has \( k \) neighbors. If every vertex is connected to every other vertex in the graph then a maximum of \( \binom{k}{2} = \frac{k(k-1)}{2} \) edges can exist between them. Let \( C^' \) be the fraction of these possible edges that actually exist between a pair of nodes. The CC is then the average of \( C^' \) over all V. The crossing process of the connections was repeated until a desired CC was obtained. The diagonal entries of the connectivity matrix representing a neuron’s connections to itself are removed.

**Algorithm 1.** Algorithm to generate connectivity matrix from ensemble learners.

1. Generate simulated time series data with realistic properties for each network.
2. Draw random chunks of sizes 500, 1000, 10,000, and 100,000 from original time series.
3. For each random chunk for each sample size:
   i. Apply discretization and conditioning.
   ii. Generate connectivity scores using base predictors—Correlation, Cross-Correlation with time lag 1, Mutual Information, Gini coefficient, and GTE.
   iii. Concatenate scores from base predictors to generate features for each pair of neurons. Training labels are obtained from ground truth data indicating presence or absence of synaptic connections.
   iv. Perform 10-fold cross-validation repeated over 10 iterations using classifiers RF, linear SVM, and NNs.
   v. Generate AUC.
4. Average AUC for each sample size for each network is calculated.

**Figure 2.** Calcium Fluorescence Histograms of Network 1 and Network 2, respectively.

2. Study design

The same preprocessing—discretization and conditioning— is applied for all analyses. In phase 1 of the study, we performed an initial benchmarking study of the algorithms performance with respect to sample size and parameters. Area under ROC curve (AUC) and Area under Precision Recall Curve (AUPR) were calculated. Ensemble learning or generating features using base predictors can improve classification [15-18]. Based on our benchmarking study, in phase 2 of the design we generated an ensemble of features. The features were constructed by concatenating scores from correlation, cross-correlation for time lag 1, mutual information, Gini coefficient, and GTE. The predictors used were Support Vector Machines (SVMs) [19], Artificial Neural networks (NNs) [20, 21], and Random Forests (RFs) [22]. Our algorithmic approach is outlined in Algorithm 1.
Linear SVMs classify into two classes, here the presence or absence of synaptic connections between a pair of neurons, by calculating the maximal-margin hyperplane separating them. We have used a LIBSVM [23] MATLAB implementation with a linear SVM and a cost parameter of 1.

RFs are an ensemble classification method which uses bagging and constructs multiple decision trees at training time. The label is the mode of the classes of individual decision trees. We have used an R implementation of the RF [24] with 100 decision trees and no pruning. We have also performed permutation testing, which revealed that results from RF were significantly better than random AUC values.

NNs are inspired by biological neural networks and can be used as pattern recognition tools. Given a specific prediction task, NNs use the input patterns to learn the class of functions by minimizing error. We have used the MATLAB Neural Network Toolbox [25] with 1 hidden layer, 10 hidden neurons, and 1000 training epochs.

We have developed a toolkit in Python consisting of algorithms for discretization, correlation-based, and information gain-based algorithms has been developed for calcium fluorescence imaging data and is available publicly [26]. Where possible this code is used. All experiments were performed on CentOS 6.3 with MATLAB v2013a and Python 2.7.3.

3. Performance and Error Estimation

We perform a 10-fold cross validation [27] and averaged results repeated over 10 iterations for two separate ensembles generated from random subsampling of the original calcium fluorescence data. The cross-validation procedure divides the subsamples drawn into 10 non-overlapping balanced subsets. The process is then repeated 10 times with 9 sets used for training and 1 for testing.

4. Performance Metrics

The network reconstruction performance was evaluated with the AUC [28]. The AUC is equivalent to the area under the curve obtained by plotting sensitivity, or true positive rate against 1 - specificity, or false positive rate at different thresholds.

Results

In phase I of our benchmarking study, for subsamples of 500, 1000, 10,000, and 100,000, the network reconstruction performance was evaluated for 1000, 500, 100, and 10 runs for conditioning levels of 0.10 and 0.15. Results were averaged over all the runs for a given network and sample size. From the Receiver Operator Characteristic (ROC) and precision recall (PR) curves we concluded that most of the reconstruction techniques converged quickly at less than 1000 time samples. Correlation-based and Information Gain-based techniques had the best reconstruction performances. We repeated the reconstruction procedure for smaller sample sizes from 100, 200, ..., 1000 for 1000 runs and AUC and AUPR with 95% confidence interval (CI) were calculated. Results were averaged over all the runs for a given network and sample size. From the ROC and PR curves we observed, that correlation-based and information gain-based algorithms had AUC ~0.8 to ~0.9 for even smaller sample sizes of 1000 and 10000. As the lags were increased, both Cross-correlation and GC performed worse. This is explained as slower time lags fail to capture the characteristic instantaneous causal interactions between neurons.

In phase II, the network reconstruction performance was evaluated for two sets of random samples of sizes 500, 1000, 10000, and 180000 drawn from the original calcium fluorescence series and repeated for conditioning levels of 0.10 and 0.15 for each base predictor and ensembles generated from them. The results from the SVM, NN, and RF classifiers repeated for 10 runs and 10-fold cross-validations on the base predictors and Ensemble 1, Network 1 are in Table 2. From Table 2, we observe both NNs and RFs boosted performance even for small sample sizes of 500 and 1000. For larger sample sizes of 180000, NNS and RFs boosted AUC performance from 0.91 to 0.94 over the best base predictors. The results from the SVM, NN, and RF classifiers repeated for 10 runs and 10-fold cross-validations on Ensemble 1, Network 2 are in Table 3. From Table 3, both NNs and RFs boosted performance even for small sample sizes of 500 and 1000. For larger sample sizes of 180000, NNS and RFs boosted AUC performance from 0.88 to 0.93 over the best base predictors. The results from the SVM, NN, and RF classifiers repeated for 10 runs and 10-fold cross-validations on Ensemble 2, Network 1 are in Table 4. From Table 4, while linear SVMs, RFs, and NNs all boost classifier performances considerably for smaller sample sizes of less than
1000, RFs and NNs improve AUC from 0.89 to 0.94 for larger sample sizes. The results from the SVM, NN, and RF classifiers repeated for 10 runs and 10-fold cross-validations on Ensemble 2, Network 2 are in Table 5. From Table 5, we observe, while linear SVMs, RFs, and NNs all boost classifier performance for even smaller sample sizes, RFs and NNs improve AUC from 0.87 to 0.93 for sample sizes of even 10,000. The ensemble learners also show convergence at approximately 1000 samples.

Table 2. AUC for Ensemble 1, Network 1

<table>
<thead>
<tr>
<th># of samples</th>
<th>Conditioning level</th>
<th>Correlation</th>
<th>Cross-correlation</th>
<th>Gini</th>
<th>MI</th>
<th>GTE</th>
<th>Linear SVM (SD)</th>
<th>RF (SD)</th>
<th>NN (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.10</td>
<td>0.83</td>
<td>0.71</td>
<td>0.83</td>
<td>0.73</td>
<td>0.61 (0.04)</td>
<td>0.78 (0.008)</td>
<td>0.88 (0.002)</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.15</td>
<td>0.86</td>
<td>0.78</td>
<td>0.86</td>
<td>0.85</td>
<td>0.61 (0.03)</td>
<td>0.86 (0.006)</td>
<td>0.86 (0.004)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.10</td>
<td>0.87</td>
<td>0.71</td>
<td>0.88</td>
<td>0.79</td>
<td>0.84 (0.005)</td>
<td>0.82 (0.004)</td>
<td>0.91 (0.002)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.15</td>
<td>0.86</td>
<td>0.79</td>
<td>0.86</td>
<td>0.85</td>
<td>0.77 (0.022)</td>
<td>0.88 (0.003)</td>
<td>0.88 (0.002)</td>
<td></td>
</tr>
<tr>
<td>10,000</td>
<td>0.10</td>
<td>0.87</td>
<td>0.87</td>
<td>0.86</td>
<td>0.86</td>
<td>0.80 (0.004)</td>
<td>0.91 (0.004)</td>
<td>0.91 (0.005)</td>
<td></td>
</tr>
<tr>
<td>10,000</td>
<td>0.15</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>0.83</td>
<td>0.91 (0.004)</td>
<td>0.93 (0.005)</td>
<td>0.93 (0.002)</td>
<td></td>
</tr>
<tr>
<td>100,000</td>
<td>0.10</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>0.86</td>
<td>0.91 (0.004)</td>
<td>0.93 (0.007)</td>
<td>0.94 (0.001)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. AUC for Ensemble 1, Network 2

<table>
<thead>
<tr>
<th># of samples</th>
<th>Conditioning level</th>
<th>Correlation</th>
<th>Cross-correlation</th>
<th>Gini</th>
<th>MI</th>
<th>GTE</th>
<th>Linear SVM (SD)</th>
<th>RF (SD)</th>
<th>NN (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.10</td>
<td>0.77</td>
<td>0.64</td>
<td>0.78</td>
<td>0.78</td>
<td>0.63 (0.009)</td>
<td>0.71 (0.006)</td>
<td>0.75 (0.02)</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.15</td>
<td>0.70</td>
<td>0.56</td>
<td>0.71</td>
<td>0.70</td>
<td>0.63 (0.015)</td>
<td>0.77 (0.006)</td>
<td>0.80 (0.008)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.10</td>
<td>0.80</td>
<td>0.68</td>
<td>0.80</td>
<td>0.79</td>
<td>0.72 (0.009)</td>
<td>0.77 (0.003)</td>
<td>0.77 (0.013)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.15</td>
<td>0.71</td>
<td>0.64</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71 (0.009)</td>
<td>0.83 (0.005)</td>
<td>0.84 (0.002)</td>
<td></td>
</tr>
<tr>
<td>10,000</td>
<td>0.10</td>
<td>0.78</td>
<td>0.74</td>
<td>0.79</td>
<td>0.78</td>
<td>0.72 (0.022)</td>
<td>0.87 (0.005)</td>
<td>0.90 (0.012)</td>
<td></td>
</tr>
<tr>
<td>10,000</td>
<td>0.15</td>
<td>0.85</td>
<td>0.82</td>
<td>0.84</td>
<td>0.83</td>
<td>0.88 (0.0066)</td>
<td>0.91 (0.003)</td>
<td>0.92 (0.003)</td>
<td></td>
</tr>
<tr>
<td>100,000</td>
<td>0.10</td>
<td>0.81</td>
<td>0.83</td>
<td>0.81</td>
<td>0.80</td>
<td>0.89 (0.0237)</td>
<td>0.91 (0.003)</td>
<td>0.92 (0.005)</td>
<td></td>
</tr>
<tr>
<td>100,000</td>
<td>0.15</td>
<td>0.86</td>
<td>0.87</td>
<td>0.84</td>
<td>0.83</td>
<td>0.88 (0.005)</td>
<td>0.92 (0.002)</td>
<td>0.93 (0.004)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. AUC for Ensemble 2, Network 1

<table>
<thead>
<tr>
<th># of samples</th>
<th>Conditioning level</th>
<th>Correlation</th>
<th>Cross-correlation</th>
<th>Gini</th>
<th>MI</th>
<th>GTE</th>
<th>Linear SVM (SD)</th>
<th>RF (SD)</th>
<th>NN (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.10</td>
<td>0.84</td>
<td>0.66</td>
<td>0.84</td>
<td>0.84</td>
<td>0.65 (0.005)</td>
<td>0.79 (0.008)</td>
<td>0.85 (0.002)</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.15</td>
<td>0.78</td>
<td>0.73</td>
<td>0.78</td>
<td>0.78</td>
<td>0.68 (0.015)</td>
<td>0.85 (0.006)</td>
<td>0.87 (0.009)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.10</td>
<td>0.86</td>
<td>0.75</td>
<td>0.86</td>
<td>0.85</td>
<td>0.74 (0.004)</td>
<td>0.82 (0.004)</td>
<td>0.86 (0.004)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.15</td>
<td>0.88</td>
<td>0.81</td>
<td>0.89</td>
<td>0.88</td>
<td>0.78 (0.004)</td>
<td>0.90 (0.004)</td>
<td>0.91 (0.002)</td>
<td></td>
</tr>
<tr>
<td>10,000</td>
<td>0.10</td>
<td>0.88</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
<td>0.86 (0.007)</td>
<td>0.92 (0.005)</td>
<td>0.91 (0.002)</td>
<td></td>
</tr>
<tr>
<td>10,000</td>
<td>0.15</td>
<td>0.89</td>
<td>0.87</td>
<td>0.89</td>
<td>0.87</td>
<td>0.91 (0.003)</td>
<td>0.93 (0.003)</td>
<td>0.93 (0.002)</td>
<td></td>
</tr>
<tr>
<td>100,000</td>
<td>0.10</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>0.85</td>
<td>0.90 (0.004)</td>
<td>0.93 (0.004)</td>
<td>0.92 (0.002)</td>
<td></td>
</tr>
<tr>
<td>100,000</td>
<td>0.15</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>0.88</td>
<td>0.91 (0.006)</td>
<td>0.94 (0.005)</td>
<td>0.94 (0.002)</td>
<td></td>
</tr>
</tbody>
</table>
Table 5. AUC for Ensemble 2, Network 2.

<table>
<thead>
<tr>
<th># of samples</th>
<th>Conditioning level</th>
<th>Correlation</th>
<th>Cross-correlation</th>
<th>Gini</th>
<th>MI</th>
<th>GTE</th>
<th>Linear SVM (SD)</th>
<th>RF (SD)</th>
<th>NN (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.10</td>
<td>0.71</td>
<td>0.59</td>
<td>0.72</td>
<td>0.71</td>
<td>0.62</td>
<td>0.79 (0.006)</td>
<td>0.72 0.004</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.15</td>
<td>0.79</td>
<td>0.67</td>
<td>0.78</td>
<td>0.76</td>
<td>0.73</td>
<td>0.84 (0.006)</td>
<td>0.80 0.008</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.10</td>
<td>0.70</td>
<td>0.57</td>
<td>0.72</td>
<td>0.71</td>
<td>0.62</td>
<td>0.84 (0.005)</td>
<td>0.80 0.007</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.15</td>
<td>0.74</td>
<td>0.68</td>
<td>0.73</td>
<td>0.72</td>
<td>0.60</td>
<td>0.80 (0.010)</td>
<td>0.85 0.006</td>
<td>0.82 0.002</td>
</tr>
<tr>
<td>10,000</td>
<td>0.10</td>
<td>0.80</td>
<td>0.72</td>
<td>0.80</td>
<td>0.80</td>
<td>0.78</td>
<td>0.79 (0.015)</td>
<td>0.90 0.002</td>
<td>0.89 0.009</td>
</tr>
<tr>
<td>10,000</td>
<td>0.15</td>
<td>0.85</td>
<td>0.82</td>
<td>0.83</td>
<td>0.82</td>
<td>0.78</td>
<td>0.88 (0.005)</td>
<td>0.91 0.002</td>
<td>0.92 0.002</td>
</tr>
<tr>
<td>100,000</td>
<td>0.10</td>
<td>0.80</td>
<td>0.82</td>
<td>0.81</td>
<td>0.80</td>
<td>0.78</td>
<td>0.89 (0.009)</td>
<td>0.91 0.004</td>
<td>0.92 0.005</td>
</tr>
<tr>
<td>100,000</td>
<td>0.15</td>
<td>0.86</td>
<td>0.87</td>
<td>0.84</td>
<td>0.82</td>
<td>0.79</td>
<td>0.88 (0.005)</td>
<td>0.92 0.003</td>
<td>0.93 0.004</td>
</tr>
</tbody>
</table>

Discussion

We performed a benchmarking study of several algorithms with different parameters based on sampling distribution and variability. Based on the performance of the algorithms, we use the scores on the connections to construct features of a classifier and predict the presence or absence of connections under different experimental conditions. While other studies have extracted topological features by combining base predictors and trained RFs on them, our study advances the state of the art by adopting a unified data-driven approach to algorithm, parameter, and feature construction combination. We have also robustly tested for overfitting by performing permutation testing.

The net result is an important enhancement of reconstruction performance from ~0.89 of the best base predictors to ~0.93 of RF and NN ensembles for large sample sizes of 100,000. This is a fully generalizable methodology applicable to all brain connectivity datasets with even more realistic numbers of neurons. The methodology also admits incorporating additional methods as they become available for further performance enhancement. For example, edge orientation algorithms can be used to infer directionality of the connections. Causal feature selection [29] on the ensemble could also be used in future expansion of the method (but was not pursued here because such methods have not yet been customized for brain connectivity). Improved signal preprocessing such as network deconvolution, spike extraction, etc. may generate further improvements on performance.

Limitations of this study include the use of simulated datasets in the absence of ground truth from real time measurements of in vivo neuronal networks. On the other hand the use of realistic simulators enables generating the necessary sample sizes, allows for determine effects of noise and a multitude of other analytic factors, while it neatly isolates assay-related experimental effects from the effects of computational analysis. When data with robust ground truth becomes available from complex in vivo organism studies the methodology presented here can be readily used to derive new discovery methods. Moreover, finer parameter selection could boost classifier performances even more. Our findings can help advance the state-of-the-art methods to design massively scalable algorithms for modeling real neuronal systems and also aid causal structure reconstruction in other areas like finance, genomics, and psychiatry where such problems are common.

In conclusion the framework presented here studies the performance characteristics and combines information theoretic, feature construction and pattern recognition based meta-learning methods to considerably improve the Area under ROC curve (AUC) performance of prior techniques. The approach is generalizable and extensible. Our data are very promising toward the feasibility of reliably reconstructing complex neuronal connectivity.

Acknowledgement

The first author would like to thank Dr. Isabelle Guyon, Javier Orlandi, Dr. Olav Stetter, Dr. Jordi Soriano, Dr. Demian Battaglia, and Dr. Mehreen Saeed for organizing the Kaggle Connectomics Challenge, providing sample code in MATLAB, data simulator, and useful discussions related to this research. The neuronal connectivity simulator and sample code in MATLAB used in some of the experiments were obtained from the Kaggle...
Connectomics Challenge [9]. The first author would also like to thank Dr. Sisi Ma for help with RF experiments and Dr. Alexander Alekseyenko, Dr. Eric Peskin, and Loren Koenig for useful feedback on the manuscript.

This work has utilized computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics at the NYU Langone Medical Center.

References

26. First Connectomics Challenge Starter Kit Python Github Repository [https://github.com/bisakha/Connectomics]
Variability in Electronic Health Record Usage and Perceptions among Specialty vs. Primary Care Physicians

Travis K. Redd, MD, MPH,1,2 Julie W. Doberne, BS,3 Daniel Lattin, MD, MS,1 Thomas R. Yackel, MD, MPH,3 Carl O. Eriksson, MD, MPH,4 Vishnu Mohan, MD, MBI,3 Jeffrey A. Gold, MD,3,5 Joan S. Ash, PhD, MLS, MBA,3 Michael F. Chiang, MD, MA1,3

Departments of1 Ophthalmology, 3 Medical Informatics & Clinical Epidemiology, 4 Pediatrics, and 5 Pulmonary and Critical Care Medicine
Oregon Health & Science University, Portland, OR
2 Department of Medicine, Legacy Medical Group, Portland, OR

Abstract
Despite federal incentives for adoption of electronic health records (EHRs), surveys have shown that EHR use is less common among specialty physicians than generalists. Concerns have been raised that current-generation EHR systems are inadequate to meet the unique information gathering needs of specialists. This study sought to identify whether information gathering needs and EHR usage patterns are different between specialists and generalists, and if so, to characterize their precise nature. We found that specialists and generalists have significantly different perceptions of which elements of the EHR are most important and how well these systems are suited to displaying clinical information. Resolution of these disparities could have implications for clinical productivity and efficiency, patient and physician satisfaction, and the ability of clinical practices to achieve Meaningful Use incentives.

Introduction
Electronic health records (EHRs) have become an increasingly critical component of modern health care delivery, and are used in all clinical disciplines.1 59% of hospitals and 48% of office-based providers currently use EHRs.2,3 However, despite the growing ubiquity of these systems, there is still substantial variability in adoption between different clinical disciplines.1,3 Specifically, adoption rates within surgical and medical specialties are approximately one half that of primary care physicians.1,4,5

In addition to these disparities in adoption rate, concerns exist that different medical fields may have varying levels of compatibility with current-generation EHRs. Numerous medical specialty societies have expressed the need for specialty-specific systems to meet the unique needs of their respective fields, including ophthalmology, orthopedic surgery, dermatology, oncology, obstetrics/gynecology, pediatrics, and pathology.6,7,8,9,10 These unique needs include differing workflow, information gathering, and clinical documentation requirements along with variations in baseline clinical volume, billing and compliance requirements, and specialty-specific terminology. Nonetheless, the widespread adoption of EHRs in the United States continues to increase, driven largely by federal incentives through the Centers for Medicare and Medicaid Services (CMS) Meaningful Use program.7,11 In 2015, penalties will begin to be levied against health care organizations that fail to meet several key EHR implementation requirements.14 Further incentivizing EHR adoption and making the avoidance of EHRs less practical for physicians, regardless of how suitable such systems are to their specialty-specific needs.

There are numerous potential implications if current-generation EHRs do not function adequately in medical disciplines of all varieties. These include decreased physician and patient satisfaction, impaired productivity and efficiency, and difficulty meeting Meaningful Use requirements.12,13 However, to date there is no experimental evidence as to whether or not such a disparity in functionality exists among different clinical disciplines, or what the nature of those differences in information-gathering needs might be. While various disciplines have expressed differing ideas of how the EHR should function and what it should provide, there is no evidence that these groups use current generation EHRs differently from each other in clinical practice.15,16 In order to better meet the health information technology needs of all clinical disciplines, determining whether such interdisciplinary differences exist and identifying their precise nature is imperative. To accomplish this, we developed a survey to characterize three parameters of physicians’ methods of clinical information gathering using EHRs when evaluating a new patient; these were 1) How the EHR is incorporated into typical clinical workflow, 2) Which elements of the chart are most important and useful to the clinician, and 3) The strengths and weaknesses of the electronic chart in displaying
relevant clinical information. These parameters were then compared between specialty and primary care physicians to identify any differences that may exist.

**Methods**

This study was approved by the Institutional Review Board at Oregon Health & Science University (Portland, OR). Acknowledgement of an information sheet by survey participants was used in lieu of informed consent.

**Survey Development**

The authors developed an 18-question survey for the purposes of data collection (Appendix). When answering these survey questions, respondents were asked to envision the scenario of evaluating a new patient rather than performing a follow up visit. This was because the former is a situation that physicians of all specialties have experience with, and because it provides the greatest insight into the strengths and weaknesses of the interaction between physician and EHR. Subjects were also instructed to envision using their own, most commonly-used EHR when responding to survey questions rather than a conception of a generic EHR system. Demographic characteristics were also collected, including primary clinical specialty, gender, clinical experience (years since graduation from medical school), level of computer experience, and primary practice setting (ambulatory vs. inpatient). The survey also included an optional free-text response elicting any additional thoughts or comments. Survey reliability was confirmed using test-retest and alternate form methods. Specifically, the survey was administered twice to 4th year medical students, each administration separated by one week. There was 91% agreement between pre- and post-test responses. Survey content and construct validity were established iteratively through expert interviews and feedback.

**Survey Administration**

An email containing a link to the questionnaire was distributed to all practicing physicians at three health care organizations in Oregon (Oregon Health & Science University/Portland VA Medical Center, PeaceHealth Medical System, and Legacy Emmanuel Medical Center) and one in Pennsylvania (Children’s Hospital of Philadelphia). These institutions were selected because of the wide variety of primary care and specialty disciplines represented at each site, and because they represented a mix of academic and community-based practices. The email was then resent to all recipients after one month. The survey was administered using REDCap electronic data capture tools hosted at Oregon Health & Science University.

**Statistical Analysis**

The primary purpose of this study was to compare several aspects of physicians’ information gathering methods using the EHR across different clinical disciplines. In order to perform this comparison, individual disciplines were combined into two groups: the Specialty group and the Primary Care group. Primary Care was considered to include General Internal Medicine, General Pediatrics, Family Medicine, and Geriatrics, in accordance with the definition of the term provided by Medicare. The Specialty group was defined as any clinical discipline other than these four Primary Care disciplines, and in this case included respondents from Obstetrics & Gynecology, Ophthalmology, Orthopedics, General Surgery, Surgical Sub-Specialties, Emergency Medicine, Internal Medicine Sub-Specialties, and Pediatric Sub-Specialties.

Three primary outcomes were compared between the Specialty vs. Primary Care groups: Outcome 1) How the EHR is incorporated into typical clinical workflow (Table 3); Outcome 2) Which elements of the chart are most important and useful to the clinician (Question 15; Figure 1); and Outcome 3) The strengths and weaknesses of the electronic chart in displaying relevant clinical information (Questions 17 and 18; Tables 4 and 5). Categorical response options were assessed using the Pearson Chi² test followed by multinomial logistic regression accounting for the covariates listed previously. Binary outcomes were assessed using the Pearson Chi² test followed by multivariable logistic regression. Ordinal outcomes were assessed using the Cochran-Armitage Test for Trend. One question (Question 15) provided multiple categorical responses for each respondent; in this case, proportions and 95% confidence intervals were compared between the Specialty vs. Primary Care groups for each potential response. Likert-type scale responses followed a nearly normal distribution and were treated as discrete continuous variables. An overall composite score was determined for each question by obtaining the group mean across all sub-sections, and these composite scores were compared between the two main predictor groups using multivariable linear regression. Thus overall scores of the ability of the EHR to display needed clinical information (Question 17) and of the severity of barriers to accessing needed information in the EHR (Question 18) were obtained from each respondent and compared between the Specialty vs. Primary Care groups. All analyses were performed using Stata SE12 (StataCorp, College Station, TX).
Results

Participant Demographics

Of the 3,649 physicians who received the survey link, 744 completed the questionnaire. This yielded a response rate of 20.4%. Of these 744 respondents, 90 were excluded either because they were not actively practicing medicine, did not use an EHR on a regular basis, or did not identify with a relevant clinical specialty, resulting in 654 responses being included in the final analysis. Three hundred fifty respondents (54%) identified with a clinical discipline in the “Specialty” group, and 304 (46%) with a discipline in the “Primary Care” group (Table 1). Subjectively, there were minimal differences between these groups with respect to clinical experience (number of years in practice), baseline computer experience, level of training, and primary practice environment (ambulatory vs. inpatient) (Table 2). However, there was a slightly higher proportion of males in the Specialty group (57%) compared to Primary Care (51%). A total of 13 EHR vendors were utilized by study participants; the most common of these were Epic (Verona, WI; 71%), Centricity (GE Healthcare, UK; 5%), CPRS/Vista (US Department of Veterans Affairs; 5%), Cerner (Kansas City, MO; 3%), and Allscripts (Chicago, IL; 3%).

Incorporation of the EHR Into Clinical Workflow (Outcome 1)

Approximately one half of physicians in both the Specialty and Primary Care groups used the EHR as the primary source of initial information when evaluating a new patient (Table 3). However, there were significant differences between the two groups with regard to the other sources of information utilized (Pearson Chi² test; p=0.02). Multinomial logistic regression confirmed this association even after adjusting for differences in level of training, amount of clinical experience, and practice setting. Specifically, Specialty physicians were significantly more likely to utilize another physician as their initial source of information on a new patient (OR=2.09, p<0.01). Primary Care physicians were significantly more likely to utilize the patient as their initial source of information than their Specialty counterparts (OR=1.47, p=0.05). There were no significant differences between the two groups regarding the likelihood of using the patient chart or a technician/ancillary staff as the primary source of initial information on a patient (Table 3).

Of the Specialty physicians surveyed, 296/332 (89%) reviewed the chart prior to entering the room with the patient, compared to 244/295 (83%) of Primary Care physicians (p=0.02). This relationship was not confounded by gender, amount of computer experience, or level of training. After adjusting for the amount of clinical experience and primary practice setting (ambulatory or inpatient), Primary Care physicians were still significantly more likely to delay chart review until during or after the patient encounter than Specialty physicians (OR=2.15, p<0.01). The duration of this initial chart review session was quite variable in both groups, with the majority of respondents

<table>
<thead>
<tr>
<th>Table 1: Clinical Disciplines Represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within a survey of 654 practicing physicians in the US</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CLINICAL DISCIPLINE</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialty</td>
<td>350 (54)</td>
</tr>
<tr>
<td>Pediatric Sub-Specialty</td>
<td>157 (24)</td>
</tr>
<tr>
<td>Internal Medicine sub-specialty</td>
<td>65 (10)</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>46 (7)</td>
</tr>
<tr>
<td>Surgical Sub-Specialty</td>
<td>26 (4)</td>
</tr>
<tr>
<td>Emergency Medicine</td>
<td>20 (3)</td>
</tr>
<tr>
<td>Obstetrics &amp; Gynecology</td>
<td>17 (3)</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>10 (2)</td>
</tr>
<tr>
<td>General Surgery</td>
<td>9 (1)</td>
</tr>
<tr>
<td>Primary Care</td>
<td>304 (46)</td>
</tr>
<tr>
<td>General Pediatrics</td>
<td>169 (26)</td>
</tr>
<tr>
<td>General Internal Medicine</td>
<td>101 (15)</td>
</tr>
<tr>
<td>Family Medicine</td>
<td>34 (5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Demographic Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of the 654 physicians participating in the survey</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>SPECIALTY</th>
<th>PRIMARY CARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Male</td>
<td>200 (57)</td>
<td>156 (51)</td>
</tr>
<tr>
<td>Female</td>
<td>150 (43)</td>
<td>148 (49)</td>
</tr>
<tr>
<td>Baseline computer experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>22 (7)</td>
<td>35 (12)</td>
</tr>
<tr>
<td>Somewhat experienced</td>
<td>262 (78)</td>
<td>193 (64)</td>
</tr>
<tr>
<td>Very experienced</td>
<td>54 (16)</td>
<td>73 (24)</td>
</tr>
<tr>
<td>Level of training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resident</td>
<td>20 (6)</td>
<td>56 (19)</td>
</tr>
<tr>
<td>Fellow</td>
<td>42 (12)</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Attending Physician</td>
<td>276 (82)</td>
<td>242 (80)</td>
</tr>
<tr>
<td>Years in practice*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-10</td>
<td>108 (32)</td>
<td>104 (35)</td>
</tr>
<tr>
<td>11-20</td>
<td>98 (29)</td>
<td>89 (30)</td>
</tr>
<tr>
<td>21-30</td>
<td>66 (20)</td>
<td>56 (19)</td>
</tr>
<tr>
<td>31-40</td>
<td>51 (15)</td>
<td>36 (12)</td>
</tr>
<tr>
<td>&gt;40</td>
<td>14 (4)</td>
<td>15 (5)</td>
</tr>
<tr>
<td>Primary practice environment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambulatory</td>
<td>154 (52)</td>
<td>158 (52)</td>
</tr>
<tr>
<td>Inpatient</td>
<td>140 (48)</td>
<td>143 (48)</td>
</tr>
</tbody>
</table>

*Self-reported years since graduation from medical school
indicating a time frame of 2-10 minutes (64% in the Specialty group and 57% in Primary Care). There was no significant difference between the two groups with respect to duration of chart review (p=0.91).

Relative Importance of EHR Elements (Outcome 2)

Participants ranked several elements of the EHR to identify the top 5 “most important” when evaluating a new patient. Specialty physicians ranked these sections as (in descending order of importance): 1: Chief Complaint, 2: Past Medical History, 3: History of Present Illness, 4: Imaging, and 5: Lab Values. Among Primary Care physicians, these sections were: 1: Medications, 2: Past Medical History, 3: Chief Complaint, 4: History of Present Illness, and 5: Problem List (Figure 1). Two individual elements of the EHR were perceived as significantly more important by the Primary Care group compared to the Specialty group; the first was the Problem List, ranked among the top 5 most important sections of the EHR by 61% of Primary Care physicians (95% confidence interval [CI]: 55-66%) compared to only 27% of Specialty physicians (95% CI: 23-32%). Secondly, the Medications section was ranked in the top 5 by 76% of Primary Care physicians (95% CI: 71-81%) compared to 44% of Specialty physicians (95% CI: 39-50%). One element of the EHR was significantly more important to Specialty physicians; this was the Imaging section, ranked in the top 5 by 50% (95% CI: 45-55%) compared to only

| Table 3: Comparison of EHR Use Practices Between Specialty and Primary Care Physicians |
|---------------------------------|-----------------|-----------------|------------------|
|                                  | SPECIALTY n (%) | PRIMARY CARE n (%) | p               |
| **Initial source of information on a new patient** |                 |                  |                  |
| Other physician (referring provider) | 53 (16)         | 23 (8)           | 0.02*            |
| Patient chart                     | 167 (50)        | 153 (52)         |                  |
| The patient                       | 98 (30)         | 109 (37)         |                  |
| Technician/Ancillary staff        | 4 (1)           | 4 (1)            |                  |
| Other                             | 9 (3)           | 7 (2)            |                  |
| **Timing of initial chart review** |                 |                  | <0.01*           |
| Before entering patient room      | 296 (89)        | 244 (83)         |                  |
| In room with patient or after exiting the room | 36 (11)        | 51 (17)          |                  |
| **Duration of initial chart review** |                 |                  | 0.91*            |
| 0-2 minutes                       | 57 (17)         | 63 (21)          |                  |
| >2-5 minutes                      | 124 (37)        | 97 (33)          |                  |
| >5-10 minutes                     | 90 (27)         | 72 (24)          |                  |
| >10 minutes                       | 61 (18)         | 63 (21)          |                  |

*Pearson Chi^2 Test; association confirmed by multinomial logistic regression adjusting for level of training, clinical experience, and inpatient vs. outpatient practice setting
*Multivariable logistic regression adjusting for clinical experience and inpatient vs. outpatient practice setting
*Cochran-Armitage Test for Trend

![figure](https://example.com/image.png)

**Figure 1: Relative Importance of Various EHR Elements among Primary Care and Specialty Physicians.**
Proportion of respondents ranking the indicated section among the top 5 “most important” EHR elements.
Hx=history; HPI=history of present illness.
27% of Primary Care physicians (95% CI: 22-32%). There were also small but statistically significant differences in the rankings of the Allergies, Social History, and Past Surgical History sections (Figure 1).

**EHR Utility and Ease-of-Use (Outcome 3)**

Two Likert-type scale questions assessed this parameter. The first (Question 17) asked respondents to rank how well information was displayed in various sections of the EHR on a scale from 1 to 5 (1 indicating the display was “Very good”, 3 indicating “neutral”, and 5 indicating “Very bad”). Average ratings of these sections ranged from mean ± standard deviation (SD) of 2.04 ± 1.08 for “Laboratory Results” to 2.64 ± 1.21 for “ICU Bedside Data” (Table 4). The composite score representing the overall ability of the EHR to display relevant clinical information had a mean ± SD of 2.40 ± 0.75 (range: 1-5). Multivariable linear regression showed no difference in this composite score between the Specialty and Primary Care groups (p=0.90). However, there was a significant association with practice setting. Specifically, ambulatory physicians rated the composite score significantly worse than inpatient physicians (2.48 vs. 2.29, respectively; p<0.01).

The second Likert-type scale question asked respondents to rank the severity of 6 potential barriers to accessing needed information in the EHR on a scale from 1 to 5 (1 indicating a “Not a barrier”, 3 indicating “Moderately strong barrier”, and 5 a “Severe barrier”). Average ratings of these barriers ranged from mean ± standard deviation (SD) of 2.84 ± 1.27 for “Information in the chart is inaccurate” to 3.32 ± 1.11 for “Others don’t record information consistently” (Table 5). The composite score of these six potential barriers had a mean ± SD of 3.11 ± 0.86 (range: 1-5). Multivariable linear regression showed a small but statistically significant difference in this composite score between the Specialty (3.20) and Primary Care (3.01) groups (p<0.01). This association was not confounded by gender, amount of computer experience, level of training, clinical experience, or practice setting.

**Discussion**

This study assessed potential differences in EHR requirements among different clinical disciplines. Key findings were: 1) Both specialty and primary care physicians relied on the EHR as the most common initial source of clinical information; 2) There were significant differences between primary care and specialty physicians regarding which sections of the EHR were considered most important; 3) Specialists identified stronger barriers than primary care physicians with regard to ability to access clinical information in the EHR.

The first key finding was that both specialists and primary care physicians identified the chart as the most important initial source of patient information. This emphasizes the critical role of EHRs in modern health care, and the potential impact of using systems that do not adequately meet all providers’ needs. Interestingly, while the importance of the EHR was uniform between both groups, its method of use and incorporation into clinical workflow were not. Specifically, primary care physicians were much more likely to delay initial chart review until during or after entering the patient room. This resulted in them being more likely to utilize the patient as their initial

---

**Table 4: Ease of Accessing Different Types of Information in the EHR**

<table>
<thead>
<tr>
<th>EHR ELEMENT</th>
<th>SPECIALTY Mean ± SD</th>
<th>PRIMARY CARE Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory Results</td>
<td>2.04 ± 1.05</td>
<td>2.04 ± 1.11</td>
</tr>
<tr>
<td>Imaging</td>
<td>2.25 ± 1.24</td>
<td>2.25 ± 1.18</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>2.27 ± 1.05</td>
<td>1.94 ± 0.96</td>
</tr>
<tr>
<td>Medication List</td>
<td>2.35 ± 1.18</td>
<td>2.23 ± 1.16</td>
</tr>
<tr>
<td>Procedure Notes</td>
<td>2.37 ± 1.04</td>
<td>2.62 ± 1.10</td>
</tr>
<tr>
<td>Operative Reports</td>
<td>2.37 ± 1.04</td>
<td>2.70 ± 1.10</td>
</tr>
<tr>
<td>History &amp; Physical Documentation</td>
<td>2.37 ± 1.34</td>
<td>2.80 ± 1.74</td>
</tr>
<tr>
<td>Outpatient Clinical Documentation</td>
<td>2.42 ± 1.14</td>
<td>2.24 ± 1.08</td>
</tr>
<tr>
<td>Discharge Summary</td>
<td>2.43 ± 1.08</td>
<td>2.28 ± 1.06</td>
</tr>
<tr>
<td>Problem List</td>
<td>2.50 ± 1.17</td>
<td>2.40 ± 1.27</td>
</tr>
<tr>
<td>Inpatient Progress Notes</td>
<td>2.51 ± 1.20</td>
<td>2.44 ± 1.22</td>
</tr>
<tr>
<td>ICU Bedside Data</td>
<td>2.65 ± 1.20</td>
<td>2.61 ± 1.22</td>
</tr>
</tbody>
</table>

SD=standard deviation

**Table 5: Severity of Six Potential Barriers to Accessing Information in the EHR**

<table>
<thead>
<tr>
<th>POTENTIAL BARRIER</th>
<th>SPECIALTY Mean ± SD</th>
<th>PRIMARY CARE Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Information in the chart is inaccurate”</td>
<td>2.94 ± 1.29</td>
<td>2.76 ± 1.24</td>
</tr>
<tr>
<td>“Information I need is not in the chart”</td>
<td>3.11 ± 1.24</td>
<td>2.98 ± 1.18</td>
</tr>
<tr>
<td>“I can’t find it in the chart”</td>
<td>3.23 ± 1.23</td>
<td>2.88 ± 1.28</td>
</tr>
<tr>
<td>“Too much information”</td>
<td>3.29 ± 1.31</td>
<td>3.40 ± 1.31</td>
</tr>
<tr>
<td>“Information is poorly displayed/difficult to interpret”</td>
<td>3.29 ± 1.20</td>
<td>3.00 ± 1.25</td>
</tr>
<tr>
<td>“Other don’t record information consistently”</td>
<td>3.38 ± 1.09</td>
<td>3.26 ± 1.13</td>
</tr>
</tbody>
</table>

SD=standard deviation
source of information, while specialists were more likely to obtain information from other/referring providers. These differences in workflow provide additional opportunities for optimization of EHRs to meet the varying needs of different disciplines.

The second key finding identified several elements of the chart that were considered important by one group but not the other. Specifically, primary care physicians showed significantly greater interest in the Problem List and Medications sections than their specialty counterparts. As one respondent stated, “I’m a surgeon…I write 2 or 3 prescriptions a month, but the patient’s pharmacy is thrust before me in almost every screen.” Conversely, specialty physicians considered the Imaging section much more important than primary care physicians. This sentiment was also echoed in the respondent comments; said one physician, “In image driven specialties, like neurosurgery, it is crucial to get actual outside imaging and not just reports. The difficulty in doing this often leads to unnecessary CT/MRI scans and better communication/transmission of these data would be valuable.”

The third key finding was that specialists face slightly stronger barriers than primary care physicians in accessing needed information from the EHR. This difference was small but statistically significant, and is consistent with the complaints raised by numerous specialty societies. One respondent summarized this by saying, “I think that most of the major systems that try to serve multiple specialties are full of an unbelievable amount of bloat. My system is specialty specific and is tailored to do exactly what I need it to do.” Said another, “The electronic medical record is very poorly organized for a pediatric ICU patient. We have to create workarounds to get the information displayed in a meaningful manner.”

These results clearly demonstrate several differences between primary care and specialty fields with respect to which elements of the EHR are considered most important when gathering clinical information, as well as their perceptions of how well these systems are able to provide such information. These differences have several important implications. The first is impaired satisfaction among physicians using systems ill-suited to their practice; one recent survey suggested that 31% of all surgical and medical specialists were “very dissatisfied” with their EHR systems, compared to only 8% of primary care providers. In addition to physician satisfaction, inefficiencies introduced by poorly-integrated EHRs could impair clinical productivity and in turn affect patient satisfaction as well. Another potential sequela of this situation is greater difficulty in achieving Meaningful Use criteria, with large potential impacts on reimbursement. This is important, as it has been shown that EHR selection is heavily influenced by financial and organizational factors independent of clinical demands. In response to this concern, several medical specialty societies have successfully advocated for the inclusion of rules, exemptions, and options in stage 2 of Meaningful Use to better suit the practices of specialists. However, prior to this study there have been no data to guide these modifications, making their adequacy uncertain. Importantly, CMS does permit Meaningful Use exclusions for providers that do not collect core measures outside their scope of practice; however, these exclusions must be applied for on an individual provider basis. This places the burden of appropriately collecting these measures on the end user rather than the system, and does not provide a large-scale solution to the problem.

The results of this study inform several potential interventions to address these concerns. First, EHRs must be targeted to meet the unique documentation needs of individual specialties. Several such “specialty-specific” systems already exist, but further assessment of the precise information-gathering requirements of each specialty is required to optimize these systems. Second, the method of implementation of EHRs across health care organizations must be carefully considered. The vast majority of EHR-using physicians in the United States practice in health systems employing a single EHR system incorporated across multiple clinical departments (the so-called “Enterprise” or “Single Vendor” EHR solution). This has benefits for interdepartmental communication and ease of logistical processes such as billing and scheduling, but as the results of this study suggest, it may be difficult for a single EHR to meet the needs of all specialties simultaneously. Alternatively, a “Best of Breed” approach involving a network of specialty-specific systems can be employed. However, establishing this network of multiple products from a variety of vendors is extremely challenging from logistical and interoperability perspectives, and can result in a fragmented and ineffectual hospital information system. More recently, a third strategy has emerged: the so-called “Best of Suite” approach. This strategy involves a point-by-point assessment of the relative merits of integration vs. differentiation at each node of the information system (i.e. individual clinical departments, billing, scheduling, etc.), resulting in a framework falling somewhere between the “Single Vendor” and “Best of Breed” models. This approach may provide a more balanced solution, improving hospital efficiency while simultaneously meeting the varying needs of different clinical disciplines as identified in this study.

This study has several limitations. First, the response rate is on the low-normal end for similar surveys of this nature. Thus our respondent pool may not be representative of the population as a whole, and may be a collection
of physicians with the most strongly-held beliefs on this topic. However, the wide ranges and standard deviations of responses to Likert-type scale questions indicate adequate variability of opinion among the respondents. Second, our grouping of clinical disciplines was fairly coarse due to overrepresentation of some disciplines compared to others. For example, there were many more pediatricians than surgeons in our respondent pool. However, subjectively there were minimal differences between individual specialties within groups, indicating an appropriate categorization scheme. Additionally, these differences resulted in inadequate power to identify differences between individual specialties, requiring the grouping of disciplines into “Specialty” and “Primary Care” categories. Consequently, our results provide a broad assessment of differences between clinical disciplines, but future studies are needed to identify differences between individual specialties. Third, not all EHR systems were represented in our study. However, our sampling scheme did capture several of the most heavily used products nationwide. Additionally, supplementary analysis revealed that the trends identified were present among both Epic and non-Epic users, indicating that the results are not unique to this system alone. Finally, Likert-type scale responses were analyzed parametrically, which assumes that the intervals between ordinal categories are of equal size. For example, we assume the difference between “Not a barrier” (1 out of 5) and “Moderate barrier” (3 out of 5) is the same as that between “Moderate barrier” and “Severe barrier” (5 out of 5). However, this assumption was supported by the fact that responses to these questions followed approximately normal distributions.

Conclusions
This study demonstrates several differences between specialty and primary care physicians in their methods of using EHRs for clinical information gathering and perceptions of the most important elements of these systems. This has important implications for clinical workflow and efficiency, patient satisfaction, physician satisfaction, and financial reimbursement. Future studies must continue to delineate the unique requirements of individual specialty fields to facilitate informed modification of EHR design, implementation, and governmental oversight.

References

2059


Acknowledgements
Research reported in this publication was supported by Oregon Clinical and Translational Research Institute (1 UL1 RR024140 01), National Library of Medicine of the National Institutes of Health (T15LM007088), Agency for Healthcare Research and Quality (1 K12 HS022981 01), and unrestricted departmental funding from Research to Prevent Blindness.

Address for Correspondence
Michael F. Chiang, MD
Departments of Ophthalmology & Medical Informatics and Clinical Epidemiology
Oregon Health & Science University
3375 SW Terwilliger Boulevard
Portland, OR 97239
Tel: 503-418-3087 | Fax: 503-494-5347 | Email: chiangm@ohsu.edu

Appendix: Selected Survey Questions

Survey Question 15:

What information is most important to you to know about a new patient? Please select the top 5.

- must provide value

☐ Chief complaint/reason for consultation
☐ History of present illness
☐ Problem list
☐ Past medical history
☐ Past surgical history
☐ Social history
☐ Family history
☐ Medication list
☐ Allergy list
☐ Laboratory values
☐ Imaging results
☐ Previous clinic note assessment & plan
☐ Other
### Survey Question 17:

In the EHR that you use, please indicate how good each section of the system is at displaying the information you need (1 = very good, 3 = neutral, 5 = very bad).

<table>
<thead>
<tr>
<th>Section</th>
<th>1, Very good</th>
<th>2</th>
<th>3, Neutral</th>
<th>4</th>
<th>5, Very bad</th>
<th>N/A, Not a feature of my EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inpatient admission note (H&amp;P)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Outpatient clinic notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Inpatient progress notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Procedure notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Operative reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Medication list</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Problem list</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Laboratory values</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Imaging results (in the EHR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Vital signs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>ICU bedside data (Ins/outs, ventilator, telemetry)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Discharge summary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
</tbody>
</table>

### Survey Question 18:

The following are potential barriers to physicians obtaining the information they need in the EHR. Please rate the severity of these barriers in preventing you from getting the information you are looking for. (1 = not a barrier, 3 = moderate barrier, 5 = severe barrier)

<table>
<thead>
<tr>
<th>Barrier</th>
<th>1, Not a barrier</th>
<th>2</th>
<th>3, Moderate barrier</th>
<th>4</th>
<th>5, Severe barrier</th>
<th>N/A, Not a feature of my EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information I need is not in the chart</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>I can't find it in the chart</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Information is poorly displayed/difficult to interpret</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Too much information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Information in the chart is inaccurate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Others don't record information consistently</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>reset</td>
</tr>
</tbody>
</table>

2062
Predicting Health Care Utilization After Behavioral Health Referral Using Natural Language Processing and Machine Learning

Nathaniel Roysden¹, Adam Wright, Ph. D¹ ¹²
¹Harvard Medical School, Boston, MA; ²Brigham and Women’s Hospital, Boston, MA

Abstract

Mental health problems are an independent predictor of increased healthcare utilization. We created random forest classifiers for predicting two outcomes following a patient’s first behavioral health encounter: decreased utilization by any amount (AUROC 0.74) and ultra-high absolute utilization (AUROC 0.88). These models may be used for clinical decision support by referring providers, to automatically detect patients who may benefit from referral, for cost management, or for risk/protection factor analysis.

Introduction

Optimization of cost while preserving or improving quality of care is a central focus of national health reform. One major predictor of high cost of treatment is mental health comorbidity—one study found that, on average, patients with a mental health diagnosis were 2.2 times more expensive than patients without a mental health diagnosis.¹ In addition to mental health diagnoses, there are other behavioral health issues that may not result in a diagnosis, but still predict increased rates of health care utilization.² This increase in utilization is theorized to be due to a decreased ability to care for self and high rates of unnecessary care.³ In addition to cost considerations, the additional health care utilization associated with behavioral health conditions may also pose the health risk of unintended negative treatment outcomes from unnecessary care.⁴

Given the potential for significant improvement of health status and reduction in health costs, patients who may benefit from behavioral health referral are an important cohort for identification. Similar to prediction of readmission, prediction of high care utilization post-referral may signal reconsideration of treatment planning, while prediction of decreased utilization may strengthen the decision to refer. A very high specificity model may even be used to automatically prompt referral for certain patients. Accordingly, the goal of this project is to be able to create a model which can accurately predict the health care cost outcomes of behavioral health referrals.

Machine Learning (sometimes also called “Data Mining”) and Natural Language Processing (“NLP”) methods have previously been used on Electronic Medical Record (“EMR”) and billing data for both predictive and text-classification purposes. While machine learning models excel at determining the content of medical notes,⁵,⁶ their application to prediction of clinical outcomes has varied. One study achieved good results predicting future morbidity of admitted cases of suspected sepsis using a neural network.⁷ Another study used a regression tree model to predict post-hospitalization suicides of Veterans Administration patients, again with good results.⁸ With regard to more specific clinical outcomes, yet another study employed machine learning models to predict cardiac arrests after an emergency department electrocardiogram.⁹ Other common models used in descriptive and predictive machine learning include support vector machines and random forests. He et al. used data mining methods to predict early hospital readmissions.¹⁰ The He et. al. paper is most comparable to this paper, due to its broad cohort and highly variable outcome. Their study differed in its focus on “feature selection and exclusive use of administrative data for ease of portability and understanding,” with use of a model that was suited to that particular approach.

We aim to use similar machine learning approaches to predict the outcomes of a patient’s first encounter with behavioral health providers (“BH patients”). The studied outcomes are cost-related in both a relative and an absolute sense. The first outcome is decreased health care utilization by any amount following a BH encounter. The second outcome is extremely high post-BH utilization, defined as utilization in the 95th or 99th percentile of non-BH patients. These predictions will be made using random forest classifiers trained on a robust clinical data set, including structured administrative data, free text from provider notes, and lab data.

Methods

Data sources
The Partners Research Patient Data Registry (“RPDR”)¹¹ is an integrated data warehouse with the ability to query an array of discrete databases storing Partners electronic health data from its various member institutions. For this study, the RPDR was queried for patients who were seen at least once in a Brigham and Women’s Hospital (“BWH”) primary
care practice. This query returned records for approximately 221,000 patients. Those records contained EMR free text notes, procedure codes (billing), encounters (scheduling), problem lists, diagnoses, lab results for A1c and Cholesterol, medication lists, and demographics. The Partners IRB approved this study.

A second set of data was downloaded from the online Medicare Physician Fee Schedule Value Files on an annual basis from 1999 to present and used to estimate standard payments for billed services.¹²

Tools
The Anaconda¹³ distribution of Python 3.4 was used with notable packages including the Python Natural Language Toolkit (“NLTK”), Matplotlib, Numpy, Scipy, and Scikit-Learn (“sklearn”), including the Scikit-Learn¹⁴ Random Forest Classifier (“RFC”)¹⁵ implementation. Other models from sklearn that were tested but not further described in this paper include AdaBoost Classifier and Random Forest Regressor.

Intervention
Clinic location was selected as the study intervention over psychiatric diagnoses because it represents the treatment benefit provided by any behavioral health services, even for patients who do not have a defined diagnosis.¹⁶ This intervention was recorded as the date of the earliest psychiatry, psychology, or behavioral health clinic encounter from the encounters dataset, based on the listed clinic location for each date of service.

Sample / Cohort Selection
Only patients with at least one behavioral health clinic visit after 2005 were included in the study cohort, reducing from 221,000 patients with at least one visit to a BWH primary care provider to 37,000 patients. Furthermore, only patients with at least one procedure code between 120 and 365 days before and after the first BH visit were included, resulting in a final cohort of 12,759 patients.

Health Care Utilization
The procedures (billing) dataset was converted from procedure codes to their most recent available Relative Value Unit (“RVU”) value via the Physician Fee Schedule files. These RVU values were paired with their corresponding dates for each patient. All entries prior to 2005 were removed from the data set, because there was a significant change in the billing data from 1995 to 2005—an exponential increase in apparent billing that likely represents changes in hospital electronic record keeping. This significantly biased early model attempts, by creating a set of patients who almost all showed increases in health care costs between 1995 and 2005. Billing data from 2005-2014 remained stable.

Health care utilization was calculated for procedure codes in the year prior and the year following the first behavioral health visit (“dateBH”) for each patient in the BH group, as shown in (Equation 1). Billing codes from more than one year prior or more than one year after the first BH visit were ignored. Patients with no billing codes more than 120 days before or without codes more than 120 days after their first BH visit were also removed. This 120 day requirement was intended to ensure that calculated values represented enduring trends in terms of usage before and after the intervention, not singular health interactions causing apparent health utilization to be inflated by a small denominator. Monthly utilization and three-month windowed utilization were each calculated in similar fashion.

\[
\text{Equation 1: } \text{Usage} = \frac{\sum_{i=0}^{n} RVU_i}{\text{dateBH} - \text{dateBH}_n} \quad \text{where date}_n \text{ is the date that maximizes the denominator}
\]
To ensure that utilization calculations were not affected by significantly different denominators, the summary statistics for the “delta days” term were calculated, as shown in (Table 1) below.

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before BH</td>
<td>302.5 ±</td>
<td>66.0</td>
<td>330.0</td>
</tr>
<tr>
<td>After BH</td>
<td>304.1 ±</td>
<td>65.7</td>
<td>331.0</td>
</tr>
<tr>
<td>Comparison</td>
<td>289.5 ±</td>
<td>68.4</td>
<td>311.0</td>
</tr>
</tbody>
</table>

Health utilization was calculated for patients without any visits to a behavioral health provider (hereafter “Comparison Group”) using the procedure codes within one year prior to the most recent recorded procedure code, also excluding patients without at least one procedure code occurring more than 120 days before the most recent procedure code. These values were used solely for reasonable cutoff values for classification of BH patients and therefore the slight difference in recorded days between BH and comparison patients does not affect the study outcome. (Table 2) below shows health care utilization in terms of RVU per day over a range of percentiles. These comparison group percentile cutoffs were then used to calculate static cutoffs for absolute utilization by the BH group. In other words, future categorization of BH patient utilization as “above the 99th percentile” means above the 99th percentile of the comparison group and does not reflect the proportion of BH patients falling into that group.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>RVU per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.125</td>
</tr>
<tr>
<td>80</td>
<td>0.379</td>
</tr>
<tr>
<td>90</td>
<td>0.733</td>
</tr>
<tr>
<td>95</td>
<td>1.251</td>
</tr>
<tr>
<td>99</td>
<td>2.666</td>
</tr>
</tbody>
</table>

**Feature Extraction**

The frequency of health contacts was calculated as the number of unique dates of service divided by the number of days between the most remote and most recent visit in the year preceding the first behavioral health visit.

The diagnoses, labs, medications, and procedures data sets each include a date with every entry; entries following the first BH visit were discarded. The demographics data set returns values accurate to its date of access; entries were recorded for patients as of the date that particular subset of the data was returned i.e. Oct-Nov, 2014, with each possible value comprising a separate binary feature in the model. Two labs, A1c and Cholesterol, were included as features according to their most recent lab value prior to behavioral health visit. Text results like “cancelled” were entered as 0. Results satisfying the regular expression for “>14.0” were entered as 14.0. Medications were included as discrete features according to their “medication code”, a structured identifier of the medication without dosage information; multiple occurrences were counted. Diagnosis and procedure codes were recorded with each unique item entered as a separate feature; their values were entered as a count of the number of individual times they were encountered in each patient’s data set.

Finally, each patient’s notes—prior to their first behavioral health visit—were concatenated and then extracted into frequency distributions (counts) of tokens, bigrams, and trigrams after removal of the NLTK English “stopwords” set, with the exception of negating terms (“no, nor”).

**Cohort Evaluation**

(Figure 2) below shows a 40x40 bin heat map of health resource utilization in RVU/day before and after the first BH clinic encounter. The likelihood of a patient having decreased health resource utilization following their first BH encounter, by any amount, was 52.8%. The plot is log scaled by number of patients in each category as specified in the color bar on the side of the image, white areas represent no data.
Figure 2: Heat map of patient utilization before/after their first behavioral health encounter

(Table 3) below displays the same information but with a coarse and nonlinear binning. Notably, while patients in the <50th percentile and 50-80 percentile groups most often remained in their cohorts, 95th and 99th percentile patients typically fell to lower percentile groups—a minority remained persistently high utilizers. This analysis also revealed a fair comparison for the “above x percentile” models—the baseline assumption that patients will continue using the same amount of health care before and after their first BH encounter.

Table 3: Percentile groups before/after BH encounter

<table>
<thead>
<tr>
<th>Percentile</th>
<th>&lt;50</th>
<th>50-80</th>
<th>80-90</th>
<th>90-95</th>
<th>95-99</th>
<th>&gt;99</th>
<th>Total Before:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;50</td>
<td>2063</td>
<td>1215</td>
<td>278</td>
<td>73</td>
<td>25</td>
<td>8</td>
<td>3662</td>
</tr>
<tr>
<td>50-80</td>
<td>1251</td>
<td>2185</td>
<td>743</td>
<td>197</td>
<td>86</td>
<td>25</td>
<td>4847</td>
</tr>
<tr>
<td>80-90</td>
<td>302</td>
<td>919</td>
<td>639</td>
<td>291</td>
<td>128</td>
<td>35</td>
<td>2314</td>
</tr>
<tr>
<td>90-95</td>
<td>97</td>
<td>324</td>
<td>333</td>
<td>230</td>
<td>179</td>
<td>49</td>
<td>1212</td>
</tr>
<tr>
<td>95-99</td>
<td>26</td>
<td>140</td>
<td>228</td>
<td>191</td>
<td>181</td>
<td>73</td>
<td>839</td>
</tr>
<tr>
<td>&gt;99</td>
<td>5</td>
<td>28</td>
<td>43</td>
<td>54</td>
<td>73</td>
<td>42</td>
<td>245</td>
</tr>
<tr>
<td>Total After:</td>
<td>3744</td>
<td>4811</td>
<td>2264</td>
<td>1036</td>
<td>672</td>
<td>232</td>
<td>12759</td>
</tr>
</tbody>
</table>

The overall effect of behavioral health visits on patient costs was quantified by converting the total amount of RVU’s counted in the before- and after- BH groups using $35.8 per RVU. At that reimbursement rate, the pre-BH group cost $78.6 Million per year, while the post-BH group cost $73.3 Million. Also of note, the 2% of pre-referral patients who were in the 99th percentile category accounted for 16% of annual pre-BH cohort costs; the 2% of patients in the post-referral group whose utilization was above the 99th percentile accounted for 17% of post-BH costs. The BH group falling into non-BH percentiles 0-80 together accounted for 24% of spending before and 28% of spending after the first BH encounter, while comprising 67% and 71% of the BH patient population, respectively.

Categorization
Three category vectors were created and used as outcomes for prediction in our models. The positive categories for each vector were: 1. Post-BH utilization greater than 99th percentile. 2. Post-BH utilization greater than 95th percentile. 3. Post-BH utilization below pre-BH utilization by any amount.

Feature Selection
The two calculated features, utilization and visit frequency, and all of the structured data were exempted from feature selection and normalization. Throughout the feature selection process, each selection method was trained on the training data and then applied to both the training and test data.

Before feature selection, token vectors, bigram vectors, and trigram vectors were separately normalized along the sample axis. Next, n-grams that occurred in fewer than 3 or more than \((n_{samples}-3)\) samples were removed from the feature set using a variance selection method. This reduced the feature space from over 17 million features to approximately 1 million. Finally, one sklearn f_classif selector, which chooses features based on their Anova F-value, was trained to select the best 100,000 features (“above percentile” model) or 20,000 features (“decreased utilization” model) from the total set of n-grams. Approximately 163,000 (83,000) features were included in each model.
Model Creation
Structured grid search was performed on a randomized test/train set to find the optimal settings and number of included n-grams for each model, detailed below. Two metrics were used to optimize grid search: Matthews Correlation Coefficient (“MCC”) (Equation 2) and F1 Score (“F1”) (Equation 3). The MCC is a metric that balances all four quadrants of the confusion matrix and is a conservative optimization approach. The F1 score reflects performance of the precision-recall curve and is less conservative than the MCC but more conservative than the Area Under the Receiver Operating Characteristic curve (AUROC).

Equation 2: \[ MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \]

TP, TN, FP, and FN are, respectively, True Positive, True Negative, False Positive, and False Negative

Equation 3: \[ F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \]

The number of trees in each forest was partially limited by the need to balance the demands of runtime and development time, especially with classifiers trained using the max_features = None argument. The number of included features was similarly limited by runtime and available memory: only about 200,000 maximum features could fit in memory when stored as a dense array.

Above Percentile Classifiers: Each Random Forest Classifier was trained using 5000 trees with samples weighted according to the inverse of class frequency and otherwise default settings.18

Increased/Decreased Utilization Classifier: 100 trees were calculated with balanced sample weights and the max_features parameter set to None, meaning that each tree would consider all possible features to determine optimal split.

Cross Validation
Stratified three-fold cross validation was performed to train and test each classifier. Feature selection methods were trained on each separate training fold and applied only to each corresponding test fold. Similarly, each classifier was created on the train fold and applied to the test fold. No information or model parameters were shared between folds of the cross-validation process. Each fold of the cross-validation required approximately 20 minutes for the “Above Percentile” classifier and approximately 40 minutes for the “Increased/Decreased” classifier. Runtime for an individual sample (patient) on a trained, loaded classifier is in the sub-seconds range.

Results
Predicting Decreased Utilization
(Figure 3) below shows the Receiver Operating Characteristic (“ROC”) and Precision-Recall (“PR”) curves corresponding to the classifier’s ability to predict that a patient’s utilization will decrease after their first BH clinic visit.

![Figure 3: ROC and PR curves for predicting decreased utilization after BH encounter.](image-url)
(Table 4) below details classifier performance metrics for each fold of the cross validation. The “Optimization” label details which performance metric was optimized to set the threshold value used in the classifier: the remaining metrics are reported at that given classifier confidence threshold. Both the Matthews Correlation Coefficient (“MCC”) and, separately, the F1 score (“F1”; also called “F-measure”) were calculated for each fold of the cross validation. With a mean Area Under the Receiver Operating Curve (“AUROC”) of 0.74 and average precision—the precision-recall curve correlate to AUROC—of 0.74-0.75, the model performs much better than random chance.

Table 4: Predicting decreased utilization after BH encounter. MCC and F1 optimized performance metrics.

<table>
<thead>
<tr>
<th>Fold</th>
<th>Optimization</th>
<th>MCC</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
<th>TPR</th>
<th>FPR</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MCC</td>
<td>0.36</td>
<td>0.70</td>
<td>0.69</td>
<td>0.72</td>
<td>0.72</td>
<td>0.36</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.33</td>
<td>0.73</td>
<td>0.62</td>
<td>0.88</td>
<td>0.88</td>
<td>0.59</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>MCC</td>
<td>0.37</td>
<td>0.72</td>
<td>0.67</td>
<td>0.77</td>
<td>0.77</td>
<td>0.42</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.34</td>
<td>0.73</td>
<td>0.62</td>
<td>0.87</td>
<td>0.87</td>
<td>0.58</td>
<td>0.36</td>
</tr>
<tr>
<td>3</td>
<td>MCC</td>
<td>0.35</td>
<td>0.72</td>
<td>0.65</td>
<td>0.82</td>
<td>0.82</td>
<td>0.49</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.33</td>
<td>0.73</td>
<td>0.62</td>
<td>0.89</td>
<td>0.89</td>
<td>0.61</td>
<td>0.37</td>
</tr>
<tr>
<td>Mean</td>
<td>MCC</td>
<td>0.36</td>
<td>0.71</td>
<td>0.67</td>
<td>0.77</td>
<td>0.77</td>
<td>0.42</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.33</td>
<td>0.73</td>
<td>0.62</td>
<td>0.88</td>
<td>0.88</td>
<td>0.59</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Prediction of Utilization Above the 99th and 95th Percentiles

(Figure 4) shows ROC and PR curves for predicting post-BH patient utilization above the 99th percentile, with corresponding performance metrics in (Table 5). (Figure 5) and (Table 6) detail the same information for prediction of utilization above the 95th percentile.

Table 5: Predicting utilization above the 99th percentile. MCC and F1 optimized performance metrics

<table>
<thead>
<tr>
<th>Fold</th>
<th>Optimization</th>
<th>MCC</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
<th>TPR</th>
<th>FPR</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MCC</td>
<td>0.288</td>
<td>0.292</td>
<td>0.358</td>
<td>0.244</td>
<td>0.244</td>
<td>0.008</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.288</td>
<td>0.292</td>
<td>0.358</td>
<td>0.244</td>
<td>0.244</td>
<td>0.008</td>
<td>0.133</td>
</tr>
<tr>
<td>2</td>
<td>MCC</td>
<td>0.252</td>
<td>0.205</td>
<td>0.121</td>
<td>0.636</td>
<td>0.636</td>
<td>0.085</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.251</td>
<td>0.245</td>
<td>0.170</td>
<td>0.429</td>
<td>0.429</td>
<td>0.039</td>
<td>0.058</td>
</tr>
<tr>
<td>3</td>
<td>MCC</td>
<td>0.260</td>
<td>0.273</td>
<td>0.240</td>
<td>0.312</td>
<td>0.312</td>
<td>0.018</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.260</td>
<td>0.273</td>
<td>0.240</td>
<td>0.312</td>
<td>0.312</td>
<td>0.018</td>
<td>0.116</td>
</tr>
<tr>
<td>Mean</td>
<td>MCC</td>
<td>0.266</td>
<td>0.257</td>
<td>0.240</td>
<td>0.397</td>
<td>0.397</td>
<td>0.037</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.266</td>
<td>0.270</td>
<td>0.256</td>
<td>0.328</td>
<td>0.328</td>
<td>0.022</td>
<td>0.102</td>
</tr>
</tbody>
</table>
The maximal values of the mean rows were used to compare the classifiers with the base case assumption that a patient would continue using the same amount of health care before and after referral. The maxima were used because, unlike said basic assumption, a classifier model can give a range of values based on the individual sample being predicted. While the 99th percentile prediction model significantly outperforms the prior utilization assumption, the 95th percentile model is only a mild improvement, with its major additional contribution being higher recall.

Table 7: Comparison of RFC metrics to constant pre- and post- BH utilization prediction metrics (“Prior”)

<table>
<thead>
<tr>
<th>Percentile Cutoff</th>
<th>Model</th>
<th>MCC</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>99%</td>
<td>RFC</td>
<td>0.266</td>
<td>0.270</td>
<td>0.256</td>
<td>0.397</td>
</tr>
<tr>
<td></td>
<td>Prior</td>
<td>0.160</td>
<td>0.176</td>
<td>0.171</td>
<td>0.181</td>
</tr>
<tr>
<td>95%</td>
<td>RFC</td>
<td>0.367</td>
<td>0.407</td>
<td>0.327</td>
<td>0.559</td>
</tr>
<tr>
<td></td>
<td>Prior</td>
<td>0.320</td>
<td>0.371</td>
<td>0.340</td>
<td>0.408</td>
</tr>
</tbody>
</table>

Discussion

Significance
This paper is novel in three ways: First, the studied intervention—a patient’s first behavioral health clinic encounter—is both novel in its scope and aligns the studied intervention with resilience theories of behavioral health; in other words, the importance of the intervention is that a patient sought and received behavioral health care, not the precise reason for that interaction. Second, the outcome of patient health care utilization as a primary endpoint, although not entirely novel by itself, has not been well studied using methods similar to this paper. Finally, in addition to structured administrative data, this model uses data derived from lab results and, less common in predictive healthcare research, free text from provider notes.

Using this robust data set and a random forest classifier, we achieved significant improvement over random chance and over prior probability for predicting patient health utilization after behavioral health referral. Furthermore, this
method achieved similar performance metrics to the latest models for predicting early hospital readmission.\textsuperscript{10} (Table 8) below compares the classifier metrics reported in He’s paper with the metrics in this paper. Ranges for He’s paper reflect their lowest and highest reported performance metrics on two different patient cohorts and applied to both same-site and outside institution data sets, as well as the class balance of their data sets in comparison to ours.

<table>
<thead>
<tr>
<th>Model</th>
<th>F-Measure</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
<th>Class Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital Readmission</td>
<td>0.19-0.37</td>
<td>0.11-0.34</td>
<td>0.47-0.85</td>
<td>0.65-0.81</td>
<td>9%-16%</td>
</tr>
<tr>
<td>Decreased Utilization</td>
<td>0.71</td>
<td>0.67</td>
<td>0.77</td>
<td>0.74</td>
<td>53%</td>
</tr>
<tr>
<td>99\textsuperscript{th} Percentile</td>
<td>0.27</td>
<td>0.26</td>
<td>0.40</td>
<td>0.88</td>
<td>2%</td>
</tr>
<tr>
<td>95\textsuperscript{th} Percentile</td>
<td>0.41</td>
<td>0.33</td>
<td>0.56</td>
<td>0.86</td>
<td>5%</td>
</tr>
</tbody>
</table>

Because of the relatively unfiltered cohort and broadly defined behavioral health intervention, the decreased utilization classifier may be able to detect patients with behavioral health problems that increase their likelihood of seeking unnecessary care, but which can be hard to for clinicians to detect or appropriately refer.\textsuperscript{20} If that is true, it may lead to reduced negative unintended treatment effects, benefiting the physical health of those patients and reducing the cost of unnecessary care.\textsuperscript{21, 22} The high-utilization classifier may also be useful in determining which patients need additional supervision, or may be useful in future work evaluating the predictors of high medical costs.

**Applications**

We envision use of these classifiers in a clinical decision support role. Such a tool would be especially useful to primary care physicians in accountable care organizations and patient centered medical homes, who have a duty to optimize both health status and costs of their panel. Thanks to the high precision across the recall range of the decreased-utilization classifier, providers could see a useful patient-specific estimation of whether referral to a behavioral health provider will be likely to reduce that patient’s health care costs. This model may also be able to detect patients who should be referred to behavioral health on an automated basis i.e. running in “the background” of the EMR. For example, patients with “confidence” values in the 0.8 to 1.0 range would have a very high chance of having decreased healthcare utilization following referral. Identification of a patient in this range could prompt an automated notification to their primary provider or to the provider of their next clinic visit, suggesting that the provider might consider whether that patient would benefit from behavioral health services. However, this should not imply that an increase in costs would not be justified by the other benefits of mental health care; additional considerations would be required in a clinical implementation setting.

**Lessons**

This project taught the student author about the use of appropriate metrics. AUROC seems to remain the standard metric for reporting classification performance in the medical literature. However, classification problems can be highly unbalanced, such as in the case of the 95th and 99th percentile models in this paper. While an AUROC approaching 0.9 may sound impressive, it does not tell the whole story. In a skewed class situation, a strong AUROC may not correlate to strong PR performance and may, if used as the main descriptive measure, mask poor performance.\textsuperscript{23} Explicitly, it is possible, as demonstrated by this study’s performance in classifying very-high utilization patients, to achieve a high ROC curve while having a low PR curve. Similarly, as in the case of the increased and decreased utilization model, while a high PR curve and strong F1 score are useful descriptors, they still ignore the influence of true negative predictions. Thus, the Matthews Correlation Coefficient, a binary classification analog of Pearson’s r that includes all four quadrants of the confusion matrix, was selected as a primary performance metric.

Sometimes, a simple model can be as powerful as a complex model. The classifiers used in this study to predict very high utilization patients were only modestly able to outperform the extremely simplistic prediction that patients will use the same amount of health care before and after the intervention. Not only is this a humbling result, but it also raises an essential question: how much can health spending be predicted by a model? Moreover, were the modest gains in model performance due to poor feature and model selection, or do they reflect the inherent randomness of the measured outcome?

**Limitations**

One significant limitation of this paper is the fact that the models were tested without a final holdout group. While parameters from model training were not shared between training phases, the core model settings were based on grid search for optimal performance on random test/train slices of the data. Therefore, although test and train groups were randomized, it is possible that the model was over-fitted to the BWH cohort, with test sets acting as internal validation, not a true, fully unseen set of test data.
This study does not have a strong comparison model or baseline clinical prediction. While some studies use clinician chart review as a baseline for comparison, this would have been cost and time prohibitive for our current study. Future research regarding the model’s ability to predict changes in cost and to detect patients who would benefit from behavioral health referral would ideally include a clinician comparison. Implementation of the model as a clinical decision support tool would also require modeled prospective implementation and true prospective implementation before being fully adopted.

A potentially significant limitation of the study is hidden in the setting. Brigham and Women’s Hospital is a world-renowned tertiary care center. The Psychiatry department at BWH, thanks to its focus on psychosomatic medicine, is particularly focused on treatment of patients with a high amount of comorbid psychiatric, behavioral, and medical complexity. Referrals for more “typical” psychiatric problems or “more healthy” patients, due to finite in-house psychiatric services, are often made to providers outside the Partners system. For these reasons, there may be a selection bias to this cohort, which may limit the generalizability of this model outside the Brigham system. Furthermore, this setting of high medical and psychiatric complexity may limit the generalizability of conclusions regarding the cost-effectiveness of behavioral health services to BWH, where this study’s data implies that behavioral health care may be mildly, but not significantly, cost saving.

Although speaking strongly to generalizability, this study used a completely unselected patient cohort and an outcome variable—costs—that can include a range of variables, including pure randomness, human behavior, and disease progression. In addition, the intervention—any sort of behavioral health clinic visit—is exceptionally broad and does not specify whether the studied patients already had mental health diagnoses or mental health treatments. This may have limited the predictive ability of the models. Additionally, it cannot be known from this study whether the intervention is causative of the observed and predicted changes, simply that there is a correlation. In other words, this model may be able to predict changes in healthcare utilization at other points in time, not just after a behavioral health visit.

**Further Work**

In order to address the limitation caused by the lack of a holdout group, a data request is currently being processed by RPDR for a set of patients from the rest of the Partners Healthcare system. This set of patients will be tested as a holdout group for the models trained on BWH patients. We are also interested in evaluating how many patients from the non-BH cohort will be classified by the model as being potential “reduced utilization” patients.

Thanks to the ability of random forest classifiers to output feature importance values, these models could be used to help discover why some patients use significantly higher levels of care after referral and could also help discover what specific features play a role in predicting beneficial outcomes of behavioral health referrals.

Now that a pipeline has been established for analysis of RPDR-sourced data, this method can be easily modified to train on and predict other outcome variables. Future work will evaluate more specific outcomes, like reduction in glycated hemoglobin (“HbA1c”) in diabetic patients following behavioral health encounters. This author also hopes that future work will be in collaboration with other Partners institutions, which may include access to summative clinical data like symptom surveys and validated risk scores; both could serve as additional interesting features and outcome variables. We also hope to utilize Partners high performance computing to annotate provider notes for the studied patients using MetaMap, which may be able to strengthen the signal gained from free text notes beyond our current “bag of n-grams” approach. This process would be a one-time computational cost that could be stored and shared with other projects.

**Conclusion**

Our study details the development of a random forest classifier that is able to predict patients who will have decreased health resource utilization following their first behavioral health visit, using structured administrative data, lab results, and free text notes as features. We are also able to identify a subset of patients who use very high amounts of resources following behavioral health referral. These models are both an improvement over comparison prior probabilities and compare favorably to similar studies. We hope that this method will be able to improve the timeliness and accuracy of referrals to behavioral health services, improving patient physical and mental health and reducing the financial costs associated with behavioral health comorbidity.
References

12. PFS Relative Value Files - Centers for Medicare & Medicaid Services [Internet]. [cited 2015 Mar 12]. Available from: http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Relative-Value-Files.html
Translational Meta-analytical Methods to Localize the Regulatory Patterns of Neurological Disorders in the Human Brain

Vanessa Sochat†*, Maude David‡, and Dennis P Wall* 

(1) Stanford Graduate Fellow, Graduate Program in Biomedical Informatics
(2) Department of Pediatrics, Systems Medicine Division
Stanford University School of Medicine
Stanford, CA 94305

Abstract

The task of mapping neurological disorders in the human brain must be informed by multiple measurements of an individual’s phenotype - neuroimaging, genomics, and behavior. We developed a novel meta-analytical approach to integrate disparate resources and generated transcriptional maps of neurological disorders in the human brain, yielding a purely computational procedure to pinpoint the brain location of transcribed genes likely to be involved in either onset or maintenance of the neurological condition.

Introduction

The literature contains a wealth of discoveries from structured functional magnetic resonance imaging (fMRI) experiments designed to better characterize the association between behavior and regions in the brain [1-8]. Advances in the field of imaging genomics have been slow due to small sample sizes, different processing strategies, and focus on one or a small set of phenotypes [9]. This has led to a landscape of isolated findings that cannot easily extend to other measures of phenotype. Pioneering work has used GWAS and higher density genomics approaches to map disease-risk variants to particular anatomical locations in the brain [10-16]. Other efforts are addressing the problem of small sample size by compiling tens of thousands of brains for study [17-25]. Here, we attempt to complement these studies with a computational approach aimed at harnessing the compendium of knowledge represented in brain imaging research published during the last two decades. Specifically, we test whether published neuroimaging research can be mined for signal relevant to understanding genetic regulation in the brain as it pertains to both behavior and to disease. We derived a set of maps between experimentally verified psychological behaviors and regional gene regulation in the normal human brain, and asked whether these regulatory maps of behavior meaningfully related to neurological impairment and psychiatric illness. This work was divided into two parts: 1) to characterize enrichment of transcriptional regulation in neurological disorders, and 2) to confirm that the genes responsible for significant enrichment are related to neurological function and impairment. This effort represents the first attempt to combine disparate data in order to map behaviors to gene regulation in specific regions of the brain and to use this mapping to identify localized brain impairment and its relationship to onset or maintenance of neurological conditions.

Methods

Our strategy focused first on integration of large-scale neuroscience results with publicly available gene expression datasets. We mined the published corpus of functional Magnetic Resonance Imaging (fMRI) to generate mappings between behavioral terms and localized regions in the brain (Section 2.1). We then triangulated these coordinates with brain-region specific patterns of gene expression to create an initial series of spatially-specific brain maps, each location having a collection of genes known to be regulated at those positions (Section 2.2). Because patterns of gene regulation in the brain contain noise, we used a statistical approach from game theory in an attempt to minimize the noise and assemble statistically significant sets of up- and down- regulated genes for each behavioral term (Section 2.3). These procedures yielded a comprehensive mapping of behaviors tested in fMRI experiments to both brain regions and genes expressed at those brain regions. To assess if these mappings could be used to inform brain regions, and concomitantly, genes impacted in neurological disorders, we tested for enrichment of the brain maps in the gene expression profiles of a collection of neurological disorders (Section 2.4). We then used
unsupervised machine learning to summarize the outcome from this enrichment experiment, and to identify regions of the brain likely to be under regulatory impairment for each disorder (Section 2.5). Finally, we confirmed through analysis of gene function that our hypothesized brain-to-disorder maps were enriched in neurological function, impairment, and previously determined to play roles in neurological disorders (Section 2.6). A summary of our workflow is included in Figure 1.

**Figure 1:** An overview of our pipeline. We summarized activation results from the neuroscience literature (Bibliome of FMRI Experiments), and generated behaviorally-relevant brain maps (Brain Maps). We used the Allen Brain Atlas to refine these maps to meaningful sets of uMp- and down-regulated gene probes (Brainterm Gene Sets), tested refined gene sets against neurological disorders (Neuro-Disease Gene Expression), and developed a novel visualization strategy to comprehensively summarize our results (Brain Lattice Visualization).

### 2.1 Associating behavioral terms to brain regions

The NeuroSynth database [26] is a robust collection of activation points from the neuroimaging literature that collects and organizes at a voxel level disparately reported data relevant to activation of brain activity and its association to behavioral terms such as “anxiety.” The NeuroSynth algorithm mines neuroimaging papers to find the coordinates of meaningful voxels provided in structured tables and generates a 2 x 2 contingency table of counts for each voxel to indicate if activation is present or absent when a specific behavior term is present or absent. A Chi-Square test of independence is then used to determine if there is a significant dependence between the term and activation, and p-values are false discovery rate (FDR)-adjusted to account for multiple hypothesis testing at a threshold of 0.05 [28]. At the time of our study, NeuroSynth contained 100,953 activation points from 3,489 neuroscience studies from the top 17 neuroscience journals. Using the procedure above, we generated brain maps of 525 psychologically relevant behavioral terms, “brainterm maps,” each voxel a binary value representing a spatial location linked to a specific behavioral term.

### 2.2 Regulatory patterns within brainterm maps

The Allen Brain Atlas (ABA) [29] contains experimentally catalogued gene expression in normal brain tissue. At the time of our study, the ABA contained expression profiles for 3,702 spatial locations in the brain from a set of six brain specimens across 58,871 gene probes. We associated each sample location with an MNI (Montreal Neurological Atlas) coordinate and cross-referenced this list with the MNI coordinates found in our brainterm maps. Because the ABA did not contain expression values for every MNI coordinate present in our brainterm maps, we adopted a sampling strategy that assigned the MNI coordinate at the center of each voxel location in our brainterm maps to the closest MNI coordinate in the ABA, allowing for a voxel center to be no more than 3mm away. We then
performed a statistical procedure to generate two lists for every braintree map, “braintem gene sets” detailed in the following section.

2.3 Refining braintree maps to regionalized gene expression in normal brain.

To ensure that our braintem gene sets associated with each of the braintem maps contained a highly robust and interrelated collection of transcripts, we deployed a strategy from game theory that has proven useful for identification of the most important genetic contributors to a condition or phenotype of interest [30]. The specific method, called Shapley Value Regression, is a coalitional game method that aims to assess the relative importance of a set of players to the outcome (in our case, a pattern of expression) of a game. The method has been described previously [31], and essentially serves to perform feature selection to derive a more precise collection of transcripts for each braintree map. A Shapley Value is calculated for each gene probe \( i \) to determine the worth of a particular set of genes, \( S \):

\[
\phi_i = \frac{1}{n!} \sum_{S} \left( P(S, i) \cup \{i\} \right) - v(P(S, i)),
\]

\[1\]

The first term calculates the Shapley value for the set of all gene probes that included gene probe, and the second term calculates the Shapley value for the set not including \( i \). \( n \) is the number of gene probes in \( N \). Intuitively, removing a gene probe that explains little or no variance will have little or no impact on the score. However, removing a gene probe that explains a large percentage of the model variance will result in a smaller Shapley Value. Since we were interested in the gene probes that were expressed relatively up and down across the entire set, we generated binary input matrices that calculated over and under expression, respectively, as the mean expression +/- 1 standard deviation to define our conditions of interest (under and over expression coded by 1, and neutral expression coded as 0). Finally, since it is computationally expensive to calculate the Shapley value according to equation 1, we used a heuristic method developed by Moretti et al [31] that is computationally tractable and has been shown to produce reliable results. We ran a resampling procedure over our observed Shapley values with the MTP function in the multtest package version 2.18 [32] to remove genes with high Shapley values that could be attributed to chance. This procedure generated 1000 bootstrap matrices, and calculated corresponding unadjusted p-values that were corrected (FDR < 0.05) to adjust for multiple comparisons. This procedure allowed us to statistically filter out genes of low or no influence, resulting in a mapping of significantly up-regulated and down-regulated genes for each of our braintem maps. We mapped the gene probes to 29,131 gene symbols to annotate the collections.

2.4 Testing the enrichment of braintem gene sets in neurological disease

We compiled a list of 466 neurological disorders and psychiatric conditions from the National Institute of Neurological Disorders and Stroke [33], and the Diagnostic and Statistical Manual (DSM) Version 5. We queried the Gene Expression Omnibus (GEO) with MedLine defined disease terms to find 173 datasets across 47 of these disorders, and downloaded these for use in our analysis [34]. We focused only on human experiments of microarray gene expression with samples from case and control, and data available to match probe identifiers to gene symbols. These experiments were loaded into R with the GEOquery package version 2.34 [35]. We manually examined each data set to assess for proper normalization using MA plot, a transformation of gene expression data onto the log intensity ratios (M) on the y-axis and the mean average scales (A) on the x-axis [36]. Proper normalization means that most gene log intensity ratios would be 0, resulting in a plot with a straight line at Y=0 [37]. In instances where the data were not properly normalized, as indicated by log intensity ratios above or below 0, we applied quantile normalization via the BioConductor package “preprocessCore” version 1.3 [38], and manually re-examined the data. Samples across each probe set were carefully filtered to include only cases and controls specific to the phenotype of interest.

We explored the sets of differentially regulated transcripts in our 47 neurological disorders using Gene Set Enrichment Analysis (GSEA). GSEA tests the hypothesis that a particular set of genes – in our case the braintem maps described above – is significantly associated with a particular phenotype. GSEA uses Kolmogorov-Smirnov statistics to score braintem maps, and then calculates significance by empirical permutation, correcting for multiple
2.5 Visually summarizing significant brainterm maps

To visually summarize sets of brainterm maps associated with brainterm gene sets enriched in a disorder, we used an unsupervised method from machine learning, Self-Organizing Maps (SOM) [40] implemented via the kohonen package in R version 2.0.18 [30] to transform our 525 brainterm maps to an intuitive 2D space, a “brain lattice.” This brain lattice is a 2D grid of nodes, each of which is labeled with one or more of the behavioral terms, and distance in the lattice corresponds to similarity of these maps. To perform the labeling, we developed a strategy to reduce a set of significant brainterm maps associated with a disorder to one summary image with the most representative voxels across the sets. We used sparse hierarchical clustering as the strategy for feature selection of this subset of voxels from the map, using the ”sparcl” package in R version 1.0.3 [41] with the squared Euclidean distance as the dissimilarity metric. We chose the tuning parameter by assessing a range of 50 values, and the optimal value was determined based on the maximal gap statistic. Sparse hierarchical clustering achieves variable selection by way of a lasso-type penalty to generate weights: a weight is calculated for each voxel such that a larger value is indicative of the region being more important for the clustering. The highest weights are indicative of the most informative features, and so to threshold our data, we converted weights to Z scores and selected those in the upper tail of the distribution (p < 0.025). The original set of brainterm maps were finally thresholded to only include these voxels, and a mean image was generated to be mapped to the brain lattice. We flattened each mean brain image associated with a disorder into a vector, and then calculated a similarity metric, the cosine distance, of each of the nodes in the SOM. A color gradient of ten levels (white to red) was generated using the RColorBrewer package in R version 1.1-2 [42] in the range of the absolute minimum value (0), to the absolute maximum value (1) across all of the disorder mean images. The final brain lattices for each disorder can then be evaluated for degree (strength of color) and extent (portion of map colored), allowing for a straightforward evaluation of the regional patterns as they pertain to groups of behavioral terms for our significant results.

2.6 Functional analysis of the significant brainterm regions and associated genes

The brain lattice provides a high level, qualitative summary of the significant brainterm maps for each result, thus it was necessary to conduct a more thorough investigation of the brain regions within these maps and to specifically examine how they compare across disorders. We used the brain structure labels (N=414) provided by the Allen Brain Atlas to assess overlap of regions with significant enrichment between disorders. We first created a matrix of disorders and regions, with each value a count of the number of times a region is present in each brainterm map with significant enrichment for the disorder. To account for the fact that some disorders have more significant brainterm maps and thus overall higher counts, we divided each disorder row by the total sum of counts for the row. We could then use the hclust function from the stats package in R version 3.0.2 [43] to perform hierarchical clustering, and generate a dendrogram to assess regional similarity between disorders. To investigate the underlying regions of the groups defined by our clustering, we took the intersection of regions shared between disorders in a group, and manually annotated each region with high level functions to do a qualitative assessment of shared function.

In order to characterize the biological functions of the gene sets associated with the diseases previously defined in this study, we mapped each gene to the KEGG (Kyoto Encyclopedia of Genes and Genomes) BRITE database [44,45]. BRITE is a collection of manually created hierarchical text files capturing the functional organization of various biological objects. It allows for the exploration of entire functional pathways, enabling us to characterize the link between disease-related genes and biological function at a variety of levels. The BRITE hierarchy is structured in 6 levels, from level A (describing biological functions at the most general levels; e.g. “Organismal Systems”), to level E (describing the precise cellular processes in which each gene is involved; e.g. “mitochondrial carrier adenine nucleotide translocator”), and finally at level F, the gene symbol. Since a gene can be involved in several pathways, we weighted each gene by dividing it by the number of different pathways in which it is involved.
Using the KEGG API, we identified the unique KEGG IDs for each gene symbol used in the GSEA. We retrieved the corresponding BRITE hierarchy and explored them using Krona, a software package that allows the exploration and visualization of BRITE data [46]. Examples of the analyses and results conducted using Krona are available in the online Supplementary materials. Additionally, we used a published tool “Genehawk” [47] to conduct PubMed queries to retrieve disorder-related, original research articles as another test of our ability to reproduce known findings about genes in disease. Genehawk is a rule-based text-mining algorithm with keyword matching that can extract target disorders, genes with significant results, and the type of study described by the article.

Results

3.1 Regulatory patterns within brainterm maps.

We generated 525 brainterm maps for a set of 525 psychologically relevant terms. Each of the maps was mapped to corresponding MNI coordinates in the Allen Brain Atlas using the procedure detailed above. Matched sets of sample points in the Allen Brain Atlas ranged in size from 27 to 952 for the 525 sets. Each of the 525 brainterm maps was associated with an up- and down-regulated set of genes, resulting in total of 1050 gene sets. Of these 1050, there were a total of 912 with a core of significant genes as determined by Shapley Value Analysis. The number of genes in these final brainterm gene sets ranged in size from 28 to 1006 genes (mean = 78.02, sd = 116.70). We systematically compared each of these sets to the pattern of gene expression in all 47 neurological disorders included in the study. Of the 47, 8 had at least 1 brainterm gene set passing the significance threshold of FDR < 0.1. These were Autism Spectrum Disorder (ASD), Alzheimer’s Disease (AD), Lupus (L), Multiple Sclerosis (MS), Parkinson’s Disease (PD), Post Traumatic Stress Disorder (PTSD), Rett Syndrome (RS), and Schizophrenia (S) (Table 1). The number of brainterm maps enriched across these 8 disorders varied from 1 to 304, with totals of 2, 304, 25, 199, 1, 1, 3, and 1 for ASD, AD, L, MS, PD, PTSD, RS, and S, respectively (Table 2).

3.2 Visually summarizing significant brainterm maps

“Brain-lattice” maps for each of our 8 disorders (Section 2.5) are available (http://www.vbmis.com/bmi/n2g). The AD and MS brain lattices showed high levels of transcriptional activity across a large percentage of brain regions (as indicated by the dark orange across a majority of the brain lattice). The brain lattice for Lupus also had a large extent of activation across the brain, but to a lesser degree, again suggesting that brainterm maps with significant genes cover a large portion of the brain. On the contrary, the brain lattices for the other disorders (ASD, PD, PTSD, SZO, and RETT), were matched to brainterm maps with more localized and far less widespread spatial activation, resulting in a smaller region of the map. For ASD, enriched regions of the brain map included those involved with emotional salience (“mood”, ”anger”, ”happy”, ”affect”, ”emotion”), as well as threat response (“threat”, ”anxiety”) and perception (“vision”). The PD brain lattice was enriched for areas related to language (“linguistic”, ”spoken”), auditory perception (“hearing”, ”sound,” ”prosody,” ”speech”), and “violation,” a term possibly indicative of “violation of social norms.” The SZO brain lattice was enriched for areas associated with synthesis of language (“comprehension,” ”sentences,” ”language,”), sound (“vocal”, ”syllable”, ”auditory”) as well as “violation.” Finally, the PTSD brain lattice was enriched for areas that correspond to decision making (“classification,” ”incorrect”), as well as emotion (“empathy”), memory (“2-back”, ”rule”), and reasoning (“non-verbal”, ”reasoning”), and RETT included a larger set of terms associated with various functions (“languages,” ”reading,” ”writing,” ”goal”, ”thinking”). Interactive brain lattices are available at http://www.vbmis.com/bmi/n2g.

3.3 Functional analysis of the significant brainterm regions and associated genes

Hierarchical clustering of our matrix of regional features described previously (Section 2.6) revealed two groups, including PD / SZO, and ASD / RETT, and the remaining four disorders (MS, AD, L, and PTSD) in isolated clusters (Figure 2). The top functions of brain regions shared by ASD and RETT included motor (globus pallidus, medial group of nuclei, red nucleus, intralaminar nuclei, lateral group of nuclei, precentral gyrus), sensory (postcentral gyrus, lateral group of nuclei, intralaminar nuclei, middle temporal gyrus, medial group of nuclei), and language (inferior frontal gyrus, supramarginal gyrus, planum polare, planum temporale, supramarginal gyrus, middle temporal gyrus). For PD and SZO, the top functions associated with the overlapping regions included language (supramarginal gyrus, planum polare, angular gyrus, middle and superior temporal gyrus), and movement (caudate...
nucleus). Overlap between PTSD, PD, and SZO included one brain region, insular cortex. Notably, the top functional annotations for regions in significant maps for ALZ and MS included cerebellum, motor, and memory, and for PD, language (supramarginal gyrus, temporal gyrus, angular gyrus), motor (caudate nucleus, putamen, substantia nigra), and relay (intralaminar nuclei, thalamus, medial nuclei, reticular nuclei, claustrum).

To confirm that the genes enriched across each of our disorders play roles in brain function and neurological impairment, we performed a functional analysis of these genes using BRITE [44,45] from the Kyoto Encyclopedia of Gene and Genomes (KEGG) database. Although each gene set contained genes that did not map to a specific function (varying from 5% of genes in AD to 51% for SZO), a large percentage of genes in each set have been determined to play roles in a small and unique constellation of pathways. The central and peripheral nervous system were enriched in 7 out of the 8 disorders (grouped and referred to as the nervous and sensory system in the BRITE category). The immune system appeared as enriched in 6 out of 8 disorders. 45 genes linked to AD were related neurological functions (Receptors, Signaling Pathways, Peripheral Nervous system and Synapses), further suggesting global brain dysregulation of gene expression. 4 of the 28 genes implicated in ASD were implicated in central nervous system functions, notably the GABRR1 gamma-aminobutyric acid (GABA) receptor subunit. The MS gene set contained 178 genes, and 7 are related to neurological functions specific to the Peripheral Nervous System. L, PTSD, and SZO affected only Receptors and Peripheral Nervous system. Finally, RETT impacted the Peripheral Nervous System via PRKCH, which is involved in inflammatory mediator regulation. Complete functional results have been made available at http://www.vbmis.com/bmi/n2g.

In addition, leveraging a PubMed literature mining tool to retrieve disorder-related original research articles [47] we determined that the gene sets associated with RETT (21 genes), ASD (26 genes) and PD (9 genes) have not been significantly linked to these diseases in the literature. 44, 7, 12, 1, and 1 of the genes associated with AD, L, MS, PTSD, and SZO, respectively, have been reported in the literature as positively associated with each disease (Table 2).

**Figure 2:** Assessment of regional similarity: We used hierarchical clustering to show groups of disorders with similar significant regions. Significant brain regions maps associated with each disorder were reduced to a set of unique anatomical labels provided by the Allen Brain Atlas, and normalized counts of these labels were clustered to show disorders with significant genes from a similar distribution of regions in the Allen Brain Atlas.

**Discussion**

In the present analysis, we used meta-analysis of functional imaging studies and a database of normal gene expression in the human brain to derive sets of up- and down-regulated genes across spatial locations, each of which was associated with a psychological term from the neuroscience literature. We then examined whether these mappings significantly associated with certain human neurological disorders, and identified significant associations
to 8 out of 47. Both the regions implicated and genes associated with the disorders have been recognized as important in neurological disease, supporting the potential of our computational approach to map neurological dysfunction to localized regions of the brain.

**Functional Assessment**

It is notable that ASD and RETT were clustered together based on regional similarity, as RETT was listed as a specific type of ASD in DSM-IV and ICD-10, and there is a large subgroup of individuals with RETT with severe symptomatology of autism [48]. Both disorders have impairment in motor, sensory, and language [49,50], and specifically, brain imaging studies have found impairment in the overlapping regions between these disorders for motor [51-54], sensory [54,55], and language [56-58]. The grouping of PD and SZO is also compelling, as both disorders are linked to malfunction of the dopaminergic system [59], and in this context, the high overlap of regions with the functional annotation “movement” supports their regional similarity. Additionally, both disorders have marked deficits in language systems [60,61], an observation that gives insight to the commonality of regions annotated with “language.” Our regional results also confirm the clinical manifestation of PD, including aberrancy in regions related to motor, relay, and language, notably the substantia nigra [62] and thalamic nuclei [63]. Similarly, for ALZ and MS, the spatial maps support the known fact that these disorders affect the brain globally [64,65].

**Limitations**

While we confirmed that brain maps of genes are implicated in neurological function, and have been previously associated with disorders (Section 3.4), we cannot claim that a gene set associated with a brainterm map indicates that the genes are directly involved with the manifestation of the behavior represented by the functional map. Our results however show promise for use of computational tools to generate a more precise understanding of brain regions, functions, and aberrant transcriptional patterns across many disorders.

**Conclusion**

We aimed to use computational approaches to map regions in the brain to neurological disease at the level of regulation of gene transcripts. We believe this is the first effort to use a purely computational approach to map the regulatory patterns of neurological disorders in the brain using meta-analysis and synthesis of published data from the field of neuroimaging genomics. Our resulting maps of transcriptional regulation in a subset of disorders is informed by spatial location and gene expression, a result that can be investigated in multiple dimensions by either zooming in to the level of genes, or out to the level of terms. The benefit of our method is that it is not limited to a particular type of brain map, type of expression profile, or method to select the gene subsets. Therefore we can use neuroimaging maps that represents regions with significantly different structure or function between phenotypes, or functional networks derived on an individual level. Additionally, the validation of the brainterm gene sets need not be limited to gene set enrichment analysis; other methods could be used to make an assessment of the signature of a set of genes in expression data. Our methods can be extended to more data types, possibly including data from different tissues, different conditions, and different cell types.

**Table 1: Summary of gene expression datasets:** Gene expression data from the Gene Expression Omnibus for each significant disorder, including the number of GEO sets, the number enriched in the disorder, and the number of cases/controls for each.

<table>
<thead>
<tr>
<th>Disorder</th>
<th># of GEO Sets</th>
<th># Enriched</th>
<th>Cases/Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autism Spectrum Disorder</td>
<td>11</td>
<td>1</td>
<td>12/14</td>
</tr>
<tr>
<td>Lupus</td>
<td>18</td>
<td>2</td>
<td>36/110; 9/17</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>14</td>
<td>2</td>
<td>45/99; 46/56</td>
</tr>
<tr>
<td>Post Traumatic Stress Disorder</td>
<td>2</td>
<td>1</td>
<td>15/12</td>
</tr>
<tr>
<td>Rett’s Syndrome</td>
<td>1</td>
<td>1</td>
<td>21/20</td>
</tr>
<tr>
<td>Disorder</td>
<td>Brainters</td>
<td>Genehawk</td>
<td>Core Genes</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------</td>
<td>----------</td>
<td>------------</td>
</tr>
<tr>
<td>Alzheimer’s Disease</td>
<td>304/0</td>
<td>1586</td>
<td>411</td>
</tr>
<tr>
<td>Autism Spectrum Disorder</td>
<td>2/0</td>
<td>826</td>
<td>26</td>
</tr>
<tr>
<td>Lupus</td>
<td>25/0</td>
<td>795</td>
<td>62</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>3/196</td>
<td>1010</td>
<td>178</td>
</tr>
<tr>
<td>Parkinson</td>
<td>1/0</td>
<td>945</td>
<td>9</td>
</tr>
<tr>
<td>Post Traumatic Stress Disorder</td>
<td>1/0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Rett Syndrome</td>
<td>3/0</td>
<td>116</td>
<td>21</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>1/0</td>
<td>1455</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 2: Brainters sets and genes associated with neuropsychiatric disorders

References

38. BM B Bolstad BM. preprocessCore: A collection of pre-processing functions. R package version 1261.
Automated Detection of Benzodiazepine Dosage in ICU Patients through a Computational Analysis of Electrocardiographic Data

Maxwell T. Spadafore¹, Zeeshan Syed, PhD¹, Ilan S. Rubinfeld, MD, MBA²
¹University of Michigan, Ann Arbor, MI; ²Henry Ford Hospital, Detroit, MI

Abstract

To enable automated maintenance of patient sedation in an intensive care unit (ICU) setting, more robust, quantitative metrics of sedation depth must be developed. In this study, we demonstrated the feasibility of a fully computational system that leverages low-quality electrocardiography (ECG) from a single lead to detect the presence of benzodiazepine sedatives in a subject’s system. Starting with features commonly examined manually by cardiologists searching for evidence of poisonings, we generalized the extraction of these features to a fully automated process. We tested the predictive power of these features using nine subjects from an intensive care clinical database. Features were found to be significantly indicative of a binary relationship between dose and ECG morphology, but we were unable to find evidence of a predictable continuous relationship. Fitting this binary relationship to a classifier, we achieved a sensitivity of 89% and a specificity of 95%.

Introduction

Maintaining proper sedation of intensive care unit (ICU) patients undergoing mechanical ventilation requires careful and continuous monitoring and adjustment of drug delivery rates. When done manually, this can be tedious, time-consuming, and more importantly, imprecise, with one study finding one preventable error for every five administered doses¹. Closed-loop computerized control systems, with their ability to integrate more information and make higher-frequency, more granular adjustments, are a promising solution to this problem, and they have already been implemented for operating room anesthesia delivery as well as postoperative hypertension management²–⁵. Their use in administration of sedatives in the ICU, however, remains largely unexplored. This may be due to the fact that, unlike anesthetic concentration or hypertension, which can be measured directly and quantitatively, sedation depth is a much more qualitative concept⁶. Traditional methods for monitoring depth of anesthesia, including the bispectral index (BIS), which is derived from electroencephalography (EEG), are inconsistent among ICU patients at the same level of sedation⁷. As a result, qualitative scales, such as the modified Ramsay sedation scale (MRSS), are often preferred⁸. Closed-loop control models have been proposed which use the MRSS as their measurable input, but they are specific to only one drug (midazolam) and make the assumption that the MRSS is always available⁹.

A more reliable quantification of sedation depth is needed. Electrocardiography (ECG) is a key biosignal with massive physiological relevance that is ubiquitous in the ICU and is often continuously monitored. As such, it is an ideal measurement source for a closed-loop control system if it demonstrates detectable changes in the presence of sedatives.

In this study we evaluated the feasibility of one aspect of this potential control system - the ability to computationally detect the presence of a given sedative (lorazepam and midazolam) using only a single lead of ECG. We present answers to three core research questions (RQs):

1. Is there a relationship between a subject’s ECG and the amount of drug in his/her system?
2. Does this relationship vary continuously with dosage, or is it binary?
3. Is the relationship (binary or continuous) strong enough to be reliably computationally predicted?

Benzodiazepines and ECG

Several studies have demonstrated the effect of lorazepam and midazolam on cardioconductive activity. Lorazepam and midazolam belong to a class of psychoactive drugs known as benzodiazepines, which function by increasing the effect of the neurotransmitter GABA. This in turn gives the drugs sedative, hypnotic, anxiolytic, anticonvulsant, and muscle relaxant properties, as well as a strong propensity for abuse¹⁰. In addition to their primary effect on neurotransmitter response, benzodiazepines also block potassium (K+) channels in cardiac tissue, which causes changes in cardiac repolarization. These changes then present themselves in the ECG as deviations in T-wave morphology, specifically as changes in T-wave length, notching, QT interval, QT variability, QT dispersion, and T-
wave area. These deviations (and others) have been shown by many to be effective in predicting K+ channel blockage, and Manini et al. demonstrated their use for successful prediction of adverse cardiovascular events in the extreme case of poisoning by benzodiazepines and other psychotropic drugs. Whereas Manini et al. used manual annotations of high-quality, hospital-grade 12-lead ECG to predict between controls and patients subjected to extremely high (toxic) doses of benzodiazepines and other drugs, our study used computational annotation of only one lead (lead II) sampled at only 125 Hz, with varying non-toxic doses of benzodiazepines.

Lorazepam and midazolam were chosen from among the pool of benzodiazepines because they are commonly prescribed for sedation in a hospital setting rather than as common anti-anxiety medication or muscle relaxants. As such, we hypothesized that lorazepam and midazolam dosage would promote significant ECG morphology changes above baseline while remaining a reasonable model for real-world benzodiazepine administration.

Methods

Data for this study were used with permission from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II clinical database, which consists of demographics, de-identified medical records, and waveforms for 32,536 patients over seven years across several individual intensive care units (ICUs) at Beth Israel Deaconess Medical Center in Boston, Massachusetts.

Subjects

Subjects were selected based on the following inclusion criteria. Subjects were given one or both of midazolam (Versed) or lorazepam (Ativan) during their ICU stay; had associated ECG waveforms, specifically lead II ECG; and experienced no cardiovascular complications of any kind during their ICU stay. The latter criterion controlled for possible cardiovascular confounders that could present at any point during a subject’s ICU stay, and was implemented by excluding any subject billed with one or more ICD9 codes between 390 and 459 (“diseases of the circulatory system”), inclusive.

Of 14 subjects who met these criteria, five were excluded due to poorly digitized (very low resolution) lead II ECG or ECG from other leads mislabeled as lead II. A total of nine subjects (eight men, one woman; mean age 42.4 years) were included in the study.

Doses

Each subject received a series of doses of midazolam or lorazepam during their ICU stay. Together, these doses formed a “dose pattern” over time that was specific to each subject (see Figure 1). Two-minute samples of lead II ECG centered on the peak effect period of each dose were acquired. Peak effect period differed for midazolam (0.5 hours after dosage) and lorazepam (3 hours after dosage). For the nine subjects, there were 539 doses (and thus 539 2-minute lead II ECG samples), with 279 as lorazepam and 260 midazolam. The dosages for the two drugs were standardized using a dose equivalency table to that of diazepam (Valium). The mean equivalent dosage was 0.253 mgkg/hr.
Figure 1. A dose pattern for a certain subject. Each subject was associated with a specific dose pattern, the plot of his/her lorazepam or midazolam dosage versus time.

For control samples, 5 minutes of lead II ECG were acquired from the moments leading up to the patient’s first dose. This served as our best approximation of the patient’s “undosed” baseline state. Because control samples were longer than dose samples, they were divided into groups of 30 beats, which were then treated as individual “control doses” for a given subject.

Beats and Feature Extraction

The PhysioToolkit programs *wqrs* and *ecgpuwave* were used to computationally label the start, peak, and end of the P-waves, QRS complexes, and T-waves within the dose and control lead II ECG samples. These annotations were further filtered so that only beats with completely annotated P-QRS-T (all of start, peak, and end for each wave) were kept. From these, the following morphological per-beat features were extracted. Length: P-wave, T-wave, QRS length, and total beat length; and Interval: QT, RT, ST, and RR.

Because dose samples were taken at many different times throughout a subject’s stay, we found it important to correct all of the above features using Bazett’s method, and we used only the corrected features in our analysis. For the remainder of this report, any reference to the above feature names implies their corrected variants.

We then extracted variability features (standard deviation) from all of the previously mentioned features.

All T-based features were extracted for their previously described connections to benzodiazepine dose. QRS length, P length, and total beat length were extracted to investigate the effect of benzodiazepines on overall beat morphology. Variability features were extracted to investigate the effect of dosage on inter-beat irregularity and autonomic nervous system regulation.
The feature extraction process (left) and data architecture (right) used in our method. Both control and dose samples undergo the same process. Beats are extracted using ecgpuwave and other PhysioNet tools, then filtered for PQRSST completeness. Length-based features and their variability-based counterparts are then extracted from the beats. Before classification, the control beats are used to normalize the dose beats, and the average for each feature across all beats is taken for each dose/control sample.

Doses as the Unit of Analysis

We chose dose/control samples, rather than beats, as the individual units of analysis, averaging the above features across all identified beats in a sample to produce one value for each feature for each sample. This decision was made for several reasons. First, doses most nearly model our target real-world situation; each dose sample represented a titration of medication. Second, although not as independent as individual subjects, doses separated by hours are far more independent than beats separated by seconds, which lends more credence to statistical analysis. Third, averaging features across all beats in a dose/control sample allowed us to reduce inter-beat variability in non-variability-based features while capturing it in the variability-based features.

Methods of Analysis

Specific methodologies were used to answer the aforementioned research questions.

To answer RQ1, we performed a two-sample, unpooled, one-sided t-test comparing the mean raw (non-normalized) values of each feature across all dose samples (at any dose level and across all subjects) against the mean raw values of each feature across all control samples. As previously described, we expected the features to increase when affected by the benzodiazepines. The number of dose samples was 536, and the number of control samples was 115. We tested at two significance levels, one without Bonferroni correction for multiplicity at \( \alpha_B = 0.05 \) and one with Bonferroni correction at \( \alpha_B = 0.05/16 = 0.0031 \). The results for these tests are given in Table 1.

In answering RQ2, we calculated the Spearman correlation coefficient \( R_s \) for each feature between each subject’s dose magnitudes and the feature, leaving us with nine correlation metrics (one corresponding to each subject) per feature. We chose the Spearman correlation coefficient over the Pearson metric because it captures both linear and nonlinear covariation between two variables and is more resistant to outliers\(^{20}\). Because the small sample size precluded the use of a parametric test of significance, we performed a 10,000-iteration permutation test on the correlation metrics for each of the features, randomizing the signs of each subject’s correlation and computing a t-statistic on each iteration. P-values were then the ratio of the number of times the randomized t-statistic exceeded the actual to the total number of iterations. The results for these tests are given in Table 2.

In our analysis of RQ3, we set up a binary classification task in an attempt to separate control samples from their dose counterparts. Features were first standardized in a by-subject fashion. Each subject’s features across each dose were standardized by the mean and standard deviation of the corresponding features from the subject’s entire control sample. Features for control samples were standardized against their own respective means and standard deviations. This method of standardization makes two key assumptions: that a subject’s baseline feature values are always known prior to their classification by the system, and that these baseline values are unbiased. We found the first assumption justifiable because pre-sedation ECG is often available and clinicians often choose to wake
their mechanically ventilated patients once per day\textsuperscript{21}. The second assumption is more tenuous in the context of our study’s data source, an issue discussed in the Limitations section.

After standardization, the data was fit to a support vector machine (SVM) with a radial basis function (RBF) kernel using scikit-learn, with parameters \( C \) and \( \gamma \) chosen by manually-narrowed grid search\textsuperscript{22}. RBF was chosen because visual inspection of features showed dose data points “curving” around controls (figure not shown). Leave-one-subject-out cross validation was used to prevent overfitting and to evaluate model performance on a by-subject basis.

Three different feature sets were used, corresponding to three different slices of the significant features taken from Table 1. In this case, significant features were taken to be those significant at \( \alpha_0 \), to allow for the maximum number of useful features to be included in the model. The first feature set included all significant features that were not variability based (5 total), while the second included all significant features that were variability based (7 total). The third consisted of all of the features from both the first and second - all significant features (12 total). For each feature set, we calculated average recall, specificity, balanced accuracy, and the area under the receiver operating curve across all subjects. Balanced accuracy is defined as one-half the sum of sensitivity and specificity and is used to correct for any imbalance between the number of dose and control samples\textsuperscript{23}. These results are presented in Table 3.

Results

RQ1: Is the computationally measured effect of non-toxic levels of benzodiazepines on ECG significant?

Table 1. Single Feature Dose vs. Control Significance

<table>
<thead>
<tr>
<th>Feature</th>
<th>( p &lt; )</th>
<th>Sig. ( \alpha_0 )</th>
<th>Sig. ( \alpha_B )</th>
<th>Feature (cont’d)</th>
<th>( p &lt; )</th>
<th>Sig. ( \alpha_0 )</th>
<th>Sig. ( \alpha_B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>P Length</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>P Length SD</td>
<td>0.01</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>T Length</td>
<td>0.002</td>
<td>Y</td>
<td>Y</td>
<td>T Length SD</td>
<td>( 1 \times 10^{-14} )</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>QRS Length</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>QRS Length SD</td>
<td>NS</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Beat Length</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>Beat Length SD</td>
<td>( 1 \times 10^{-5} )</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>QT Interval</td>
<td>0.02</td>
<td>Y</td>
<td>N</td>
<td>QT Interval SD</td>
<td>( 1 \times 10^{-9} )</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>RT Interval</td>
<td>0.003</td>
<td>Y</td>
<td>Y</td>
<td>RT Interval SD</td>
<td>( 1 \times 10^{-13} )</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>ST Interval</td>
<td>( 1 \times 10^{-7} )</td>
<td>Y</td>
<td>Y</td>
<td>ST Interval SD</td>
<td>( 1 \times 10^{-10} )</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>RR Interval</td>
<td>0.001</td>
<td>Y</td>
<td>Y</td>
<td>RR Interval SD</td>
<td>0.04</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

Evident in Table 1, the benzodiazepines exerted a significantly large effect on the features. As expected based on the previous work described in the “Benzodiazepines and ECG” section, T-based features showed large differences, with QT interval significant at \( \alpha_0 \) and T length and ST and RT intervals significant at \( \alpha_B \). This shows that ECG is indeed affected by benzodiazepines even at moderate doses, and that computational methods can be relied upon for the extraction of features traditionally extracted manually by cardiologists from a high-quality 12-lead ECG.

Of interest and warranting further investigation are the small \( p \)-values resulting from the variability-based features. An increase in variability arising for the doses could be caused by a variety of factors, with some legitimate, including physiological compensation for dosage and differences in dose patterns, and some confounding, such as wavering subject condition.

RQ2: Is this effect binary or continuous?

With the analysis from the previous section showing that there was indeed a computationally identifiable effect of benzodiazepines on ECG, we investigated if the magnitude of this effect varied with the magnitude of the dosage.
Table 2. Single Feature Dose-Control Spearman Correlation

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean $R_S$</th>
<th>$p &lt;$</th>
<th>Sig. $a_0$</th>
<th>Sig. $a_B$</th>
<th>Feature (cont'd)</th>
<th>Mean $R_S$</th>
<th>$p &lt;$</th>
<th>Sig. $a_0$</th>
<th>Sig. $a_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P Length</td>
<td>-0.22</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>P Length SD</td>
<td>0.16</td>
<td>NS</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>T Length</td>
<td>0.14</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>T Length SD</td>
<td>0.24</td>
<td>NS</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>QRS Length</td>
<td>0.00</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>QRS Length SD</td>
<td>-0.02</td>
<td>NS</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Beat Length</td>
<td>0.02</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>Beat Length SD</td>
<td>0.28</td>
<td>NS</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>QT Interval</td>
<td>0.12</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>QT Interval SD</td>
<td>0.30</td>
<td>NS</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>RT Interval</td>
<td>0.11</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>RT Interval SD</td>
<td>0.32</td>
<td>NS</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>ST Interval</td>
<td>0.11</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>ST Interval SD</td>
<td>0.33</td>
<td>0.02</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>RR Interval</td>
<td>0.12</td>
<td>NS</td>
<td>N</td>
<td>N</td>
<td>RR Interval SD</td>
<td>0.28</td>
<td>0.03</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

Only two features, ST Interval SD and RR Interval SD, showed a significant Spearman correlation with dosing, and their respective significant correlation coefficients were quite low. Simple visual inspection (see Figure 3) demonstrated that even these significant relationships were tenuous and would yield poor predictability. As such, it was decided that our data did not support the existence of a simple continuous relationship between benzodiazepine dosage and our features.

However, physiological reactions to drugs are extremely complex, and our model - a time-delayed linear relationship - was extremely simple, though two of our features showed significance even when this simplified model was employed. A more robust model of dose interactions through time, paired with our features, could elucidate a clearer, more predictable continuous relationship between ECG and benzodiazepine dosage.

![Figure 3.](image)

Figure 3. Although the Spearman correlation $R_S$ was significant at $a_0 = 0.05$ for feature RR Length SD, visual inspection showed a general lack of predictable correlation, illustrated here.
RQ3: Is this effect reliably predictable?

Table 3. Classification Scores

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Specificity</th>
<th>Balanced Acc.</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-SD</td>
<td>0.86</td>
<td>0.85</td>
<td>0.86</td>
<td>0.94</td>
</tr>
<tr>
<td>Only SD</td>
<td>0.82</td>
<td>0.65</td>
<td>0.74</td>
<td>0.71</td>
</tr>
<tr>
<td>Both</td>
<td>0.89</td>
<td>0.95</td>
<td>0.92</td>
<td>0.97</td>
</tr>
</tbody>
</table>

With a maximum sensitivity of 89%, a maximum specificity of 95%, and a maximum AUC score of 0.97, our method successfully differentiated between each subject’s control and dose samples. This demonstrates that a completely computational, unaided system is capable of recognizing the presence of sedative benzodiazepines using only a single lead of ECG.

For our best-performing feature set, although recall and specificity are at their highest overall, the recall is a full six points lower than the specificity. This provides some proof that some subjects may have been under the influence of benzodiazepines prior to their entry into the ICU. With untainted controls, more separation between dose and control features could be attained, leading to increases in both sensitivity and specificity.

Figure 4. Receiver Operating Characteristic curve for the Non-SD + SD feature set

Discussion and Conclusion

Summary

Due to the qualitative nature of sedation, automated sedative administration in an ICU setting via closed-loop control systems is largely unexplored. We envision a system that uses ECG to quantitatively monitor a patient’s physiological response to sedatives, and we have demonstrated the feasibility of the drug-detecting portion of this system, focusing specifically on benzodiazepines. We started with features commonly examined by cardiologists looking for evidence of cardiotoxicity in poisonings, and we generalized the annotation and extraction of these features from manual annotation of high-quality 12-lead ECG to automated annotation of low-quality ECG from a single lead. We tested the predictive power of these features using nine subjects from the MIMIC II clinical database.
who had received lorazepam or midazolam during their stay, controlling insofar as possible for any pre-existing cardiovascular conditions. Features were found to be significantly indicative of a binary relationship between dose and ECG morphology, but our simplified model for dose interactions was unable to find evidence of a predictable continuous relationship. Fitting this binary relationship to a classifier, we were able to detect the influence of benzodiazepines with a sensitivity of 89% and a specificity of 95%.

Limitations
This study has limitations. With this proof-of-concept study, we were aiming for hypothesis generation and demonstration of preliminary feasibility. As such, we traded sample size for optimal experimental conditions, namely the lack of cardiovascular complications during the subjects’ ICU stays and the need for Lead II ECG to be recorded for the subject in the MIMIC II database (only 2,809 of the 32,536 subjects had associated waveforms). These two restrictions alone reduced the subject pool from 2,971 patients to the nine our study included. As we progress past this proof-of-concept, we will remove these restrictions, increasing our sample size.

Previous medical history for each subject, including possible cardiovascular issues, was located in textual dictations of their attending physicians. As such, it was not possible for us to exclude subjects with a prior history of heart disease. However, these subjects were excluded if they were billed for any treatment related to their cardiovascular conditions at any time during their ICU stay. In a similar vein, information regarding the events of the subject’s hospital stay leading to transfer to the ICU was unavailable, and thus a control sample could have been tainted if the patient had received benzodiazepines during their stay prior to the ICU. There is some evidence suggesting that this may have been the case. This confounder would bias the results against any relationship between dosage and ECG morphology, however, so we do not see it as invalidating this study’s results. Additionally, we were unable to control for change in some extraneous factors, such as pain, intubation, and oxygenation, between control and dose periods.

For the sake of ensuring a reasonable sample size, patients were not excluded if they died while in the hospital. However, only one subject died while in the ICU, and this was verified not to affect the results. While our results are promising, our small sample size limits the external generalizability of our findings.

Future Directions
The high accuracy of our method warrants further exploration and reproduction in larger samples. To demonstrate the validity of our findings, a larger observational study, with less restrictions on eligible subjects, should be conducted. Additionally, a more powerful model for benzodiazepine dose interactions through time should be investigated to aid in the prediction of dose as a continuous, rather than binary, variable.

Acknowledgments
MTS wishes to thank Professor Ben Hansen of the Department of Statistics for his advice regarding statistical analysis of RQ1 and RQ2.

References
COBE: A Conjunctive Ontology Browser and Explorer for Visualizing SNOMED CT Fragments

Mengmeng Sun¹, Wei Zhu¹, Shiqiang Tao¹,², Licong Cui¹,², PhD, Guo-Qiang Zhang¹,², PhD
¹Department of EECS, Case Western Reserve University, Cleveland, OH
²Division of Medical Informatics, Case Western Reserve University, Cleveland, OH

Abstract

Ontology search interfaces can benefit from the latest information retrieval advances. This paper introduces a Conjunctive Ontology Browser and Explorer (COBE) for searching and exploring SNOMED CT concepts and visualizing SNOMED CT fragments. COBE combines navigational exploration (NE) with direct lookup (DL) as two complementary modes for finding specific SNOMED CT concepts. The NE mode allows a user to interactively and incrementally narrow down (hence conjunctive) the search space by adding word stems, one at a time. Such word stems serve as attribute constraints, or “attributes” in Formal Concept Analysis, which allows the user to navigate to specific SNOMED CT concept clusters. The DL mode represents the common search mechanism by using a collection of key words, as well as concept identifiers. With respect to the DL mode, evaluation against manually created reference standard showed that COBE attains an example-based precision of 0.958, recall of 0.917, and $F_1$ measure of 0.875. With respect to the NE mode, COBE leverages 28,371 concepts in non-lattice fragments to construct the stem cloud. With merely 9.37% of the total SNOMED CT stem cloud, our navigational exploration mode covers 98.97% of the entire concept collection.

Introduction

Search and browsing interfaces are an integral part of ontological system dissemination. Although ontology search interfaces are distinct from general search interfaces because of the semantic and structural information already contained in the ontologies, they can still benefit from the latest developments in information retrieval. For example, conjunctive exploratory navigation interfaces CENI and SCENI [1, 2, 3] have been developed for exploring consumer health questions with health topics as dynamically search tags complementing keyword-based lookup. The conjunctive exploration mechanism allows users to quickly narrow down to the most relevant results in the most effective way.

In this paper, we introduce a conjunctive ontology browser and explorer (COBE) for searching SNOMED CT concepts and visualizing SNOMED CT fragments. The COBE search interface has three prominent features. First, it supports both direct lookup and navigational exploration to retrieve concepts. In the direct lookup mode, a user comes with specific terms or SNOMED CT identifier, enters into the search interface, and retrieves a list of relevant concepts. In the navigational exploration mode, a user may not have a targeted term, or cannot easily or effectively formulate descriptive lookup terms, and may rely on navigational features to browse and explore the concepts. COBE provides conjunctive search combining direct lookup with navigational exploration. Second, COBE meets the need of a concept retrieval interface to reach and visualize erroneous SNOMED CT fragments for quality assurance. Third, in the navigational exploration mode, COBE utilizes a cloud of core stems mined from non-lattice fragments. In previous study [4], mining SNOMED CT non-lattice fragments have been shown to be an effective approach for detecting abnormal structures that are inconsistent with the principle that the subsumption relation (is-a) in an ontological system should conform to the lattice property. This principle is further enforced in more recent work [5], where non-lattice fragments have shown up to 40 times the change rate against background changes in the evolution of SNOMED CT.

The effectiveness of COBE for retrieving SNOMED CT concepts is evaluated in two ways. For the direct lookup mode, it is compared with the well-known IHTSDO SNOMED CT Browser [6] and NLM SNOMED CT Browser [7]. Evaluation against manually created reference standard showed that COBE attains an example-based precision of 0.958, recall of 0.917, and $F_1$ measure of 0.875. For the navigational exploration, we conduct an evaluation experiment on the search stems. The result shows that only using stems generated from 28,371 concepts which occurring in the upper bounds (uppermost level) of fragments, which is 9.37% of total, as navigational exploration menus covers 98.97% of all SNOMED CT concepts.

2092
1 Background

1.1 SNOMED CT

SNOMED CT is the world’s largest and most comprehensive clinical healthcare terminology, developed and maintained by the International Health Terminology Standard Development Organization (IHTSDO). From a structural perspective, SNOMED CT can be seen as a series of large directed acyclic graphs, one for each of its 19 sub-hierarchies: Procedure, Physical force, Event, Staging and scales, Substance, Environment or geographical location, Situation with explicit context, Body structure, Observable entity, Pharmaceutical/biologic product, Physical object, Qualifier value, Special concept, Specimen, Social context, Clinical finding, Organism, Linkage concept, and Record artifact. Each concept comes with a unique SNOMED CT identifier, and concept descriptions associated with it. The version of SNOMED CT used in this study is dated September 1, 2014. It contains 302,902 active concepts, which are organized into 19 hierarchies linked by relationships such as is-a.

1.2 SNOMED CT Browsers

SNOMED CT Browsers are applications and tools used for viewing terminology content or hierarchy structure. Typically, a SNOMED CT browser can display descriptions and relationships of a concept by providing searching terms or identifier. The official IHTSDO SNOMED CT Browser[6] is an online, multilingual, multi-edition ontology browsing application. It enables browsing of International Edition of SNOMED CT, as well as various National Editions. User can enter terms or navigate in the taxonomy hierarchies to find target concept, with filters in semantic, refset, or language. The application can provide almost everything related to the concept, including descriptions, parents and children, attributes and diagram of structure. Another online browser, provided by U.S. National Library of Medicine (NLM), is NLM SNOMED CT Browser[7]. Different from the IHTSDO SNOMED CT Browser, searching and displaying SNOMED CT content, as included in UMLS Metathesaurus, is provided in NLM SNOMED CT Browser. It expands searches to include synonymous terms from over 100 Metathesaurus vocabulary sources. CSIRO provides a graph-based interactive interface for navigating SNOMED CT concepts, which is called Shrimp[8]. One can expand the hierarchy graph by clicking the concept, which provides an intuitive visualization of its relationships.

1.3 Conjunctive Navigational Exploration

There are two basic search modes in information retrieval. One is direct lookup (DL), and the other is navigational exploration (NE). Direct lookup is a basic search mode where a user knows precisely what to look for and comes up with input strings leading uniquely to the search target. On the other hand, in the navigational exploration mode, a user may not be able to easily and effectively formulate a descriptive search string, and must rely on navigational mechanisms such as menus to browse and explore the content in order to inform the user “what is there.”

Conjunctive Exploratory Navigation Interface (CENI[1]) is a recently introduced technique for effective access of online consumer health information using navigational exploration. CENI was achieved by assigning multiple topics for information items using semantic tagging[2, 3]. Crowd sourced comparative evaluation revealed that anonymous users from Amazon Mechanical Turk preferred 2 to 1 for CENI against other search mechanisms[1].

A defining feature of CENI is an interface which provides the user multiple paths to quickly narrow down to relevant contents. By selecting one topic at a time, in an incremental fashion, a user arrives at narrower and narrower content areas that are relevant to all the topics selected so far, conjunctively.

A novel feature of COBE is the NE mode, adapted from CENI. Instead of consumer health topics as menus in CENI, COBE uses word stems from an important sub-collection of SNOMED concepts – those appearing in non-lattice fragments as defined in[4]. Such word stems serve as attribute constraints, or “attributes” in Formal Concept Analysis, which allows the user to navigate to specific SNOMED CT concept clusters in multiple ways.

1.4 SNOMED CT Quality Assurance

It has been noted that the biomedical domain is developing tremendously. For example, SNOMED CT releases a new version almost every 6 months. The quality of an ontology is a key issue that determines its usability. Many auditing methods exist to ensure the quality of ontologies. They include but are not limited to lexical[9], structural[10],
semantic [11, 12] and statistical-based [13]. A lattice structure is often one of the criteria of the well-formedness of ontologies [4]. A lattice is a structure that every two concepts in the ontology have no more than one minimal common ancestor or maximal common descendant. MaPLE [5] has found a large number of such non-lattice structure in SNOMED CT. Figure 1 is a non-lattice fragment from SNOMED CT. The double-circled concepts “Tissue specimen from breast” and “Tissue specimen from heart” share two minimal ancestors: “Tissue specimen” and “Specimen from trunk,” highlighted in pink, which makes them a non-lattice fragment. To make it lattice-conforming, one could add the concept “Tissue specimen from trunk” (dashed in Figure 1) [4, 5, 10].

Figure 1: A non-lattice fragment in SNOMED CT [5].

2 Methods

Figure 2 depicts the overall architecture of the proposed COBE to search SNOMED CT concepts and facilitate the retrieval of non-lattice fragments in SNOMED CT. For the navigational exploration mode, SNOMED CT concepts are preprocessed (dotted rectangular box) to obtain a collection of core stems for tagging concepts. SNOMED CT concepts are filtered by non-lattice fragments, and performed word tokenization and frequency ranking. As a result, a collection of core stems are mined and used for tagging concepts. In the navigational exploration (NE) mode of the COBE search interface, a user’s input (NE-input in Figure 2) is the selection of stem tags, based on which COBE performs NE-based conjunctive search and returns to the user a list of relevant concepts as well as numbers of related non-lattice fragments. In the direct lookup (DL) mode of the COBE search interface, a user’s input (DL-input in Figure 2) is a specified term formulated by the user. COBE performs term splitting, and DL-based conjunctive search to find a list of relevant concepts as well as numbers of related non-lattice fragments and returns them to the user.

2.1 Navigational Conjunctive Exploration

In the navigational exploration mode, a user may want to explore “what is there,” or may not be able to easily or effectively formulate a descriptive lookup term to find non-lattice fragments. In either case, the user may rely on navigational menus or facets to browse and explore. To address such need, we construct a cloud of informative terms serving as the navigational menus for the user to effectively explore SNOMED CT concepts and non-lattice fragments. To achieve this, we perform the following preprocessing steps:

• First, we extract all active concepts from the SNOMED CT distribution files. For each active concept, its identifier and descriptions (including fully specified name, preferred name, and synonyms) are further extracted and stored in a MySQL database.

• Second, we filter active concepts by non-lattice fragments, and only keep those concepts appearing in the uppermost level of non-lattice fragments. If a concept appears in the uppermost level of a non-lattice fragment, then the non-lattice fragment is called related to the concept.

• Third, for each of the remaining concepts, we remove punctuation marks and all stop-words such as “of” and “out” from its descriptions, which is then tokenized into individual stems. The resulting collection of tokenized singular words forms our COBE stem collection. Note that this is different from “word stems” in linguistics which refer to the prefix strings serving as the root of a word. When we mention stems in this paper, it is meant to be in the sense of COBE stem.

• Fourth, for each stem, we compute the frequency it appears in the remaining concepts to determine the order of the stems in the search menus.
After the above steps, the stems ranked by the precomputed frequencies serve as the navigational menus. Each stem is guaranteed to hit at least one fragment, which prevents the circumstance that retrieved concepts from COBE have no related non-lattice fragment.

Given a selection of input stems by a user, COBE performs conjunctive search to retrieve matching concepts, that is, only concepts matching all the selected stems are returned, ranked by the number of related non-lattice fragments in the descending order. COBE allows concepts being investigated and narrowed down by several times of navigation sequentially. For example, clicking navigation stem “heart” followed by clicking “attack” will narrow down the results to “Myocardial infarction (disorder).” COBE enables the combination of navigational exploration and direct lookup. For instance, clicking “attack” followed by typing “heart” in the direct search box will narrow down the results to “Myocardial infarction (disorder).”

2.2 Direct Lookup

The other way to navigate to non-lattice fragments is specifying a targeted term in the direct search box instead of formulating a term by the available navigation stems. In this direct lookup mode, a user knows what to look for, comes with specific term or exact SNOMED Identifier, and tries to retrieve a list of concepts. After a user types a specific term into the direct search box, COBE splits the term into stems and performs conjunctive search to retrieve concepts matching all these stems, and returns them by the number of related non-lattice fragments in the descending order.

2.3 Evaluation

We evaluate COBE in two ways. For the navigational exploration mode, we evaluate if the search stems obtained from the preprocessing step covers a wider range of concepts. To calculate the percentage of the coverage, we use the total number of concepts as the denominator, and the number of concepts that can be reached by at least one search stem as
For the direct lookup mode, we evaluate the search performance of the COBE interface by comparing it to other two SNOMED CT browsers: IHTSDO SNOMED CT Browser [6] and NLM SNOMED CT Browser [7]. We randomly select 24 concepts as search tasks (one lookup term per task) from the CORE Problem List Subset of SNOMED CT [14] to compare the three browsers. Since most SNOMED CT concepts have short length of descriptions, the terms chosen for the search tasks are 1-word, 2-word, 3-word or 4-word to make sure a certain number of concepts get returned by the browsers. 6 concepts (search tasks) are randomly selected for each word length. For each search task, an evaluator uses three browsers to retrieve a list of relevant concepts, respectively. A reference standard is created for each search task: the common concepts found by all three browsers are considered correct results and included in the reference standard; for the other concepts found but not shared by all three browsers, the evaluator manually reviews them and includes the relevant concepts into the reference standard only if the evaluator considers the result concepts correct. We use the example-based precision, recall and $F_1$ measure [2, 3, 15, 16] to evaluate the performances of the three browsers.

Let $R$ be the reference standard consisting of $m = 24$ search tasks $\{(s_i, Y_i) \mid i = 1, \ldots, m\}$, where $Y_i$ is the set of all concepts included in the reference standard for the search task $s_i$. Let $Z_i$ be the set of concepts retrieved from a search interface for $s_i$. The example-based precision ($P$), recall ($R$) and $F_1$ measure ($F_1$) are calculated as follows:

$$P = \frac{1}{m} \sum_{i=1}^{m} \frac{|Y_i \cap Z_i|}{|Z_i|}, \quad R = \frac{1}{m} \sum_{i=1}^{m} \frac{|Y_i \cap Z_i|}{|Y_i|}, \quad F_1 = \frac{1}{m} \sum_{i=1}^{m} \frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|}$$

(1)

### 3 Results

Among all the 302,902 active concepts, only 28,371 are in uppermost level of non-lattice fragments. We used 28,371 concepts to generate 12,623 search stems for navigational exploration. Table 1 displays the 10 most frequent stems appearing in the concepts related to non-lattice fragments.

<table>
<thead>
<tr>
<th>Search Stem</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>structure</td>
<td>2,640</td>
</tr>
<tr>
<td>finding</td>
<td>1,292</td>
</tr>
<tr>
<td>disorder</td>
<td>1,070</td>
</tr>
<tr>
<td>neoplasm</td>
<td>1,041</td>
</tr>
<tr>
<td>procedure</td>
<td>971</td>
</tr>
<tr>
<td>joint</td>
<td>755</td>
</tr>
<tr>
<td>artery</td>
<td>654</td>
</tr>
<tr>
<td>bone</td>
<td>641</td>
</tr>
<tr>
<td>system</td>
<td>629</td>
</tr>
<tr>
<td>entire</td>
<td>615</td>
</tr>
</tbody>
</table>

Table 1: Top-10 most frequent navigation search stems and their frequencies. Frequency: the number of concepts involved in non-lattice fragments a stem appears in.

#### 3.1 Conjunctive Ontology Browser to Explore SNOMED CT Fragments (COBE)

Figure 3 shows a sample screenshot after clicking “neck” and “head” from search stems for navigational exploration and typing “region” for direct search. The left column displays the list of navigation terms, while clicking the navigation menu bars of the left column, the chosen words are displayed inside the horizontal bar on the top of the right column, where the “Reset” button is used to start a new exploration by clearing the search words. The leftmost button of the right column shows the number of non-lattice fragments where the upper bounds contain this concept, where green button indicates existing non-lattice fragments and orange for no fragments found. The center area indicates the fully specified name and blue button for the SNOMED CT concept identifier. When hover the name COBE shows all the preferred terms and synonyms. Clicking either the green button or the concept name links to the page for visualizing relevant fragments. By default, all the concepts are displayed in the descending order of the number of fragments, if neither search term nor search keyword is specified.

#### 3.2 Visualizing Non-lattice Fragments

As shown in the right column in Figure 3, COBE enables user to find some specified concepts listed in the concept displaying area. Once the concepts are displayed, a user can click the leftmost green or orange button to link to the index page of related fragments and finally direct users to the fragments visualization page.
Figure 3: A screenshot of the conjunctive ontology browser explorer interface COBE.

Figure 4: Searching “neck head region” in both NE and DL modes to render related non-lattice fragments. The upper left part of Figure 4 demonstrates the navigational exploration mode of COBE, and upper right part for the direct lookup. Using each mode or combining using these two modes leads to the visualization part. The lower left part of Figure 4 is the fragment index page for the concept “Pain of head and neck (disorder).” The left column of the index page contains all fragments denoted by upper bounds concept identifiers, and right column displays the corresponding concept labels. After clicking the concept identifiers of a fragment, the browser directs to the page for...
the visualization of that fragment which is shown in the lower right part. The green nodes represent the upper bounds of the fragment, the yellow ones represent lower bounds and the gray ones represent intermediate nodes in between. The solid gray edges represent is-a relationship of two concepts. For example, “Chronic pain (finding)” and “Pain of head and neck region (finding)” are upper bounds of the fragment, while “Chronic neck pain (finding)” and “Chronic pain in face (finding)” are lower bounds of the fragment. The leftmost gray edge means “Chronic neck pain (finding)” is-a “Chronic pain (finding).”

3.3 Navigational Exploration Evaluation

Instead of using descriptions of all 302,902 concepts, we used a subset of 28,371 concepts related to fragments to generate the search stems for navigational exploration. Each of the 28,371 concepts must occur in the upper bounds of a fragment at least one time. 12,623 search stems were generated after punctuation marks removed, tokenized, and stop-words removed. To evaluate the search stems, we queried all search stems for each of the 302,902 concepts, and found that 299,789 concepts can be reached through at least one the search stems. This implies that using navigational exploration, at most 3,113 concepts might be missed. These 3,113 concepts must have no occurrence in the upper bounds of fragments. The coverage of the search stems to reach overall concepts is 98.97% (299,789/302,902). This indicates that our search stems formed by a small subset of concepts can cover most SNOMED CT concepts.

3.4 Direct Lookup Evaluation

To evaluate the search interface, we formulated 24 search tasks to compare the three SNOMED CT browsers’ performances: IHTSDO SNOMED CT Browser, NLM SNOMED CT Browser, and COBE interface. We randomly picked 24 concepts from the CORE Problem List Subset of SNOMED CT and used their descriptions as the searching key words for three browsers. To make sure certain number of results can be returned, we limited the search tasks to 1-word, 2-word, 3-word or 4-word concepts. Then the evaluator used three browsers to perform the search tasks. The reference standard was created for each search task: the common concepts found by all three browsers are considered correct results and included in the reference standard. For those concepts found but not shared by all three browsers, the evaluator manually reviewed them and decided whether to include them into the reference standard or not, according to the relevancy of the result and the search task. Only those concepts considered relevant by the evaluator were included into the reference standard.

Table 2 shows the search tasks, the number of concepts in reference standard, the number of concepts found by each browser and the number of concepts included in reference standard found by each search interface. Three interfaces yielded the same results for 17 of the 24 search tasks. The example-based precision ($P$), recall ($R$) and $F_1$ measure ($F_1$) were calculated based on the results and the reference standard using formula (1) in Section 2.3. Table 3 shows the overall example-based precision, recall, and $F_1$ for the results of 24 search tasks using NLM, IHTSDO and COBE. The result shows COBE carried out the best recall of 0.917; NLM achieved the best precision of 1.0; both COBE and NLM had the best $F_1$ measure of 0.875. This indicates that COBE is comparable to the state-of-the-art SNOMED CT browsers for looking up concepts.

4 Discussions

Our evaluation of the direct lookup mode of COBE is limited with respect to the reference standard. Since SNOMED CT contains 302,902 concepts, going through the entire concept list to find all relevant concepts of each search task to build the reference standard is infeasible for the evaluator. The other limitation is that we only had one evaluator. This also could possibly introduce bias to the reference standard creation. A third limitation is that only stem coverage was used to evaluate the navigational exploration mode of COBE. It would be helpful to conduct a comparative user evaluation for the navigational exploration mode. However, there are no similar mechanisms in other ontology browser for us to compare COBE with.

We also performed an experiment using the entire SNOMED CT to generate search stems and compare the performance with the search stems from concepts involved in non-lattice fragments. 83,139 search stems were generated from the entire 302,902 concepts. The number of stems generated by non-lattice fragments was 12,623, which is 15.18% of the entire stems. However, such 15.18% of the entire stems already covers 98.97% of SNOMED CT concepts.
Table 2: Comparison of searching results of 3 SNOMED CT Browser. $Y$ - the number of concepts in reference standard $Y$. $Z_C$ - the number of concepts found by COBE. $Z_C \cap Y$ - the number of concepts found by COBE that included in reference standard. $Z_I$ - the number of concepts found by IHTSDO browser. $Z_I \cap Y$ - the number of concepts found by IHTSDO browser that included in reference standard. $Z_N$ - the number of concepts found by NLM browser. $Z_N \cap Y$ - the number of concepts found by NLM browser that included in reference standard.

<table>
<thead>
<tr>
<th>Search Term</th>
<th>$Y$</th>
<th>$Z_C$</th>
<th>$Z_C \cap Y$</th>
<th>$Z_I$</th>
<th>$Z_I \cap Y$</th>
<th>$Z_N$</th>
<th>$Z_N \cap Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panniculitis</td>
<td>39</td>
<td>39</td>
<td>39 (39)</td>
<td>39</td>
<td>39 (39)</td>
<td>39</td>
<td>39 (39)</td>
</tr>
<tr>
<td>Hepatoblastoma</td>
<td>2</td>
<td>2</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Neurilemmoma</td>
<td>13</td>
<td>13</td>
<td>13 (13)</td>
<td>13</td>
<td>13 (13)</td>
<td>13</td>
<td>13 (13)</td>
</tr>
<tr>
<td>Oligomenorrhea</td>
<td>3</td>
<td>3</td>
<td>3 (3)</td>
<td>3</td>
<td>3 (3)</td>
<td>3</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Thalassemia</td>
<td>64</td>
<td>64 (64)</td>
<td>57 (57)</td>
<td>63</td>
<td>63 (63)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hand pain</td>
<td>2</td>
<td>2</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Mononucleosis syndrome</td>
<td>5</td>
<td>5</td>
<td>5 (5)</td>
<td>5</td>
<td>5 (5)</td>
<td>5</td>
<td>5 (5)</td>
</tr>
<tr>
<td>Deficiency anemias</td>
<td>2</td>
<td>2</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Viral screening</td>
<td>4</td>
<td>4 (4)</td>
<td>4 (4)</td>
<td>4</td>
<td>4 (4)</td>
<td>4</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Angina decubitus</td>
<td>1</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Renal hypertension</td>
<td>17</td>
<td>17 (17)</td>
<td>16 (16)</td>
<td>16</td>
<td>16 (16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep terror disorder</td>
<td>1</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Mantle cell lymphoma</td>
<td>4</td>
<td>4 (4)</td>
<td>4 (4)</td>
<td>4</td>
<td>4 (4)</td>
<td>4</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Tricuspid incompetence, non-rheumatic</td>
<td>1</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Acute frontal sinusitis</td>
<td>2</td>
<td>2</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Renal tubular acidosis</td>
<td>11</td>
<td>11 (11)</td>
<td>11 (11)</td>
<td>11</td>
<td>11 (11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iatrogenic Cushing’s disease</td>
<td>1</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Drusen of optic disc</td>
<td>4</td>
<td>4 (4)</td>
<td>4 (4)</td>
<td>4</td>
<td>4 (4)</td>
<td>4</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Therapeutic drug monitoring assay</td>
<td>1</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Malignant neoplasm of brain</td>
<td>14</td>
<td>14 (14)</td>
<td>13 (13)</td>
<td>12</td>
<td>12 (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bronchopulmonary dysplasia of newborn</td>
<td>1</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Antenatal ultrasound scan abnormal</td>
<td>1</td>
<td>2 (1)</td>
<td>2 (1)</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Derangement of temporomandibular joint</td>
<td>2</td>
<td>1</td>
<td>1 (1)</td>
<td>2</td>
<td>2 (2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The example-based precision, recall, and $F_1$ measure for the results of concept search using COBE, IHTSDO and NLM.

<table>
<thead>
<tr>
<th>SNOMED CT Browser</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>COBE</td>
<td>0.958</td>
<td>0.917</td>
<td>0.875</td>
</tr>
<tr>
<td>IHTSDO</td>
<td>0.958</td>
<td>0.792</td>
<td>0.75</td>
</tr>
<tr>
<td>NLM</td>
<td>1.0</td>
<td>0.875</td>
<td>0.875</td>
</tr>
</tbody>
</table>

In related work, existing ontology visualization toolkits such as KC-Viz [17] and ProtégéVOWL [18] are not designed for searching and visualizing non-lattice fragments. Therefore, COBE uniquely fulfills this need.

5 Conclusion

In this paper, we presented a conjunctive ontology browser COBE to search and explore SNOMED CT concepts and non-lattice fragments for ontology quality assurance. The direct lookup and navigational exploration of COBE allow multiple entry points for users to explore information of interest. The combination of searching and visualization mechanisms in COBE provides a novel way for structural auditing of SNOMED CT. Navigational exploration and direct lookup offer complementary modes for finding specific SNOMED CT concepts. With respect to the direct lookup mode, evaluation against manually created reference standard showed that COBE attains an example-based precision of 0.958, recall of 0.917, and $F_1$ measure of 0.875. With respect to the navigational exploration mode, COBE leverages 28,371 concepts in non-lattice fragments to construct the stem cloud. With merely 9.37% of the total
SNOMED CT stem cloud, our navigational exploration mode covers 98.97% of the entire concept collection. COBE has been deployed as the resident search interface for linking, exploring, and rendering of SNOMED CT non-lattice fragments, which represent an active area of ontology curation work.

References


Improving guideline concordance in multidisciplinary teams: preliminary results of a cluster-randomized trial evaluating the effect of a web-based audit and feedback intervention with outreach visits

Mariëtte M. van Engen-Verheul, MSc1, Wouter T. Gude, MSc1, Sabine N. van der Veer, MSc, PhD2,4, Hareld M.C. Kemps, MD, PhD1,3, Monique M.W. Jaspers, MSc, PhD1, Nicolette F. de Keizer, MSc, PhD1, Niels Peek, MSc, PhD4

1Department of Medical Informatics, Academic Medical Center/ University of Amsterdam, Amsterdam, The Netherlands; 2European Renal Best Practice (ERBP) Methods Support Team, University Hospital Ghent, Ghent, Belgium; 3Department of Cardiology, Máxima Medical Centre, Veldhoven, The Netherlands; 4MRC Health e-Research Centre, Institute of Population Health, University of Manchester, Manchester, UK

Abstract

Despite their widespread use, audit and feedback (A&F) interventions show variable effectiveness on improving professional performance. Based on known facilitators of successful A&F interventions, we developed a web-based A&F intervention with indicator-based performance feedback, benchmark information, action planning and outreach visits. The goal of the intervention was to engage with multidisciplinary teams to overcome barriers to guideline concordance and to improve overall team performance in the field of cardiac rehabilitation (CR). To assess its effectiveness we conducted a cluster-randomized trial in 18 CR clinics (14,847 patients) already working with computerized decision support (CDS). Our preliminary results showed no increase in concordance with guideline recommendations regarding prescription of CR therapies. Future analyses will investigate whether our intervention did improve team performance on other quality indicators.

Introduction

The widespread uptake of electronic patient records (EPRs) provides unprecedented opportunities to monitor clinical performance and improve care quality. As such, audit and feedback (A&F) interventions based on interrogating EPR databases are increasingly used to aid health care professionals in improving their performance. A&F provide care professionals with an objective summary of their clinical performance over a specified period of time [1]. Yet despite their expanded use, A&F interventions show variable effectiveness on improving quality of care. A recent Cochrane review of 140 randomized trials of A&F interventions reported a median 4.3% absolute improvement (interquartile range 0.5% to 16%) in quality of care, with a quarter of the studies showing a strong positive effect, but with another quarter showing a negative or null effect [1].

Previous studies have attributed much of the observed variability in effect of A&F interventions to feedback design characteristics and contextual factors. They suggested A&F to be most effective if baseline performance is low, when feedback is provided by a supervisor or colleague, more than once, both verbally and in writing, and when it includes explicit targets and an action plan [1-4]. Furthermore, the effect of indicator-based performance feedback is likely to be stronger when it is combined with educational meetings [1]. Other suggested effect modifiers are the perceived quality of the data underlying the feedback, motivation and interest of the recipient, organizational support for quality improvement (QI), and how performance targets or benchmarks are derived [5]. Besides these literature results we had a conjunctural expectancy based on the actuality that modern medicine, including the care for chronically ill patients, is not just a matter of individual professionals but largely the responsibility of multidisciplinary teams embedded in complex organizations. Therefore we expected that, specifically in chronic disease management, engaging the entire multidisciplinary teams and their managers in the QI process is an important success factor of A&F interventions [6]. We developed a multifaceted A&F intervention that both incorporates successful characteristics described in the literature [1-5] and that is specifically directed at multidisciplinary teams [7]. It comprises the use of a web-based system that provides periodic performance feedback with benchmark comparisons and support for concrete QI action planning. In combination with educational outreach visits the systems facilitates active team engagement in improving their performance [8].

We implemented our intervention in the field of cardiac rehabilitation (CR) in the Netherlands. Within this field an EPR system with computerized decision support (CDS) functionalities was previously developed to stimulate
concordance with guideline recommendations for the patient tailored CR program [9]. Although the CDS has proven to be effective [10], there remained considerable non-concordance due to a lack of resources and other organizational constraints [11]. Further improvement of guideline concordance required organizational and procedural changes that individual users considered to be beyond their own tasks, influence and responsibilities. Use of the CDS system alone was insufficient for inciting users to involve decision makers at team and organizational level to realize those changes [11]. This finding stressed the need for an intervention specifically directed at the decision-making processes at these levels to create the necessary conditions and resources for further improving guideline concordance. Hence, guideline concordance was one of the targeted behaviors of our A&F intervention.

We performed a multicenter cluster randomized trial to assess the effect of a multifaceted A&F intervention on clinical performance of multidisciplinary teams in the field of CR. In this paper we present preliminary results regarding the intervention’s effect on concordance of CR therapies with guideline recommendations.

**Methods**

**Setting: cardiac rehabilitation and computerized decision support**

CR is a therapy provided by multidisciplinary care teams to support cardiac patients recovering from a cardiac incident or intervention on both the physical and psychosocial domain [12, 13]. CR is recommended for all patients who have been hospitalized for an acute coronary syndrome (ACS) and for those who have undergone a cardiac intervention [12, 14]. A meta-analysis shows consistent evidence of the effectiveness of exercise-based and multimodal (e.g., psychosocial and stress management) CR interventions with regard to mortality and prevention of future cardiac events (relative-risk reduction 21–47%) [15]. CR teams usually include cardiologists, physical therapists, nurses, psychologists, dieticians, social workers, and rehabilitation physicians. However, in many Western countries, CR services are under-utilized and poorly standardized and do not follow the available scientific evidence [16]. A recent study in the Netherlands shows that only a minority of patients eligible for CR actually receive it [17]. The CR uptake rate was 28.5% among patients with an ACS and/or intervention.

Consistent with international standards, the Dutch Guidelines for CR [18] state that professionals should conduct an extensive needs assessment procedure (NAP) where 80 to 130 data items concerning the patient’s medical, physical, psychological, and social condition and lifestyle are gathered. Based on this procedure, a patient-tailored rehabilitation program should be prescribed which can contain up to four group-based therapies: two psychosocial therapies (disease-specific education; lifestyle modification) and two physical therapies (exercise training; relaxation and stress management training), and can be supplemented by individual counselling (e.g., by a psychologist, dietician or social worker) when needed. In the Netherlands, there are two commercial vendors of EPR systems with CDS for CR (referred to as EPR1 and EPR2) that can be used for data collection. Both systems are based on the Dutch guidelines for CR [18] and follow the same data model. They guide their users through the NAP and provide advice for the decision about each out of the four considered CR therapies for the prescribed CR program [19]. However, one of the EPRs is a stand-alone product in which, based on results of usability evaluation of a beta version of the system [20], the data entry navigational structure is organized flexible around an overview screen. Complete data collection is stimulated by showing users which steps of the NAP they already have finished and which steps they still need to complete. The other system is integrated into the hospital EPR from one vendor and offers a more straightforward data entry structure. This system does not provide feedback on finished NAP steps. After data collection in both systems, the patient specific CDS advice is discussed with the patient. Thereafter the prescribed CR program, including the decisions for each out of the four CR therapies (which can deviate from the CDS advice), are recorded in the EPR. While we focus on prescribed therapies in this study, we note that there are sometimes discrepancies between prescriptions and therapies that are actually received by patients. After participation in the program (which typically lasts for 6-12 weeks), patient are reassessed to determine results.

**Study design**

The effect of the intervention was evaluated in a multicenter cluster-randomized study in which each CR clinic received the A&F intervention, but its contents were randomly limited to one of two complementary domains that jointly constitute CR: the psychosocial domain (disease-specific education; lifestyle modification) or the physical domain (exercise training; relaxation and stress management training). In this way, both study arms served as each other’s control, and we minimized the risk of clinics dropping out of the study because they did not receive any intervention. Cluster-randomization was chosen to avoid contamination among professionals within the same clinic. We refer to the study protocol for further details of the experimental design [7].
Eligible CR clinics and patients

All CR clinics that used an EPR with CDS during the NAP and that were willing to share their data for research and to set up a local QI team were eligible to participate in the study. There were 91 CR clinics in the Netherlands, the majority affiliated with hospitals [21]. Twelve clinics were located in specialized rehabilitation clinics [21], which have regional functions and can treat both simple and more complex referred patients. All types of clinics were eligible to participate in the study, provided that they worked with either one of two commercial EPR systems for CR that could be used for data collection for our study. During the inclusion period of the trial from July 2012 until December 2013, this was the case for 22 clinics. The study dataset consisted of (i) patient identification data (31 items), (ii) CR needs assessment data (80–130 items), (iii) data on selected rehabilitation goals and therapies (79 items), and (iv) CR evaluation data (105 items). All consecutive CR patients that underwent the NAP in one of the participating clinics during the study period were eligible for enrollment in the study. Clinics that participated agreed to enter all data of these patients in their EPR.

Intervention

Our multifaceted A&F intervention was provided through a web-based system, called CARDSS Online [8]. When designing the intervention we followed the A&F literature which underlines the importance of combining periodic performance feedback with benchmark comparisons, action planning with concrete, self-formulated goals, and educational outreach visits to actively involve care professionals in a continuous QI process. To this end CARDSS Online supports four tasks: (i) monitoring of indicator-based performance by means of quarterly feedback reports including benchmark information, (ii) selecting indicators for QI that are locally perceived as important and upon which improvement is deemed feasible, (iii) developing a QI plan consisting of QI goals and concrete actions to accomplish these goals, and (iv) during follow-up iterations updating the QI plan based on new performance measurements and experiences with executing the QI actions in practice. Benchmark comparisons were summarized by a colored icon next to each indicator score which depicted whether the performance was acceptable (green checkmark), borderline (yellow checkmark), or poor (red exclamation mark). The benchmark comparisons were based on the clinic’s performance score and the average score across all clinics (details available in [7]).

Educational outreach visits were held following each feedback report. During these visits CARDSS Online was used to guide the clinics’ local QI teams through the process of systematically defining, implementing and monitoring QI actions. The QI teams consisted of at least one local CR coordinator (usually a specialized nurse), one professional from another discipline (e.g. a physical therapist), their manager and the responsible cardiologist. The visits were chaired by an investigator (MvEV) who supported the QI team with interpretation of the feedback and drafting (or during follow-up visits monitoring and updating) a concrete QI plan. Indicators were included in the plan based on the benchmark information and discussion on importance, feasibility, and expected time needed to improve. For each quality indicator, the QI team could specify the problem, presumed causes, improvement goal, and concrete actions on how to reach that goal. If clinics agreed upon extended participation after the minimum study period of one year (comprising of four A&F iterations), they received up to two more quarterly feedback reports in combination with telephone support rather than a face-to-face visit.

We previously developed a set of eighteen primary quality indicators to provide performance feedback in our system [22]. This was done in close collaboration with an expert panel (representatives from all disciplines involved in CR) and patient panel, using a modified RAND method [22]. Results from both panels were combined with results from a literature search and guideline review in an extensive rating and consensus procedure. The expert panel did not select concordance of prescribed therapies with the guidelines as one of the eighteen primary quality indicators. However, the tailoring of CR programs to individual needs of patients is an important quality theme in CR, and indirectly reflected by many indicators that were chosen. Furthermore, we did include concordance of prescribed therapies with the guidelines in the feedback reports, which enabled QI teams to include improvement actions aimed at guideline concordance in their QI plans. Besides results on indicators and concordance, the feedback also included patient characteristics (e.g. age and diagnosis), information referring to general processes (e.g. time between discharge and NAP) and structures (e.g. presence of patient satisfactory research) to reduce the risk of attrition.

Outcome measurement

Our intervention was targeted at health care professionals and was therefore expected to have a direct effect on process outcomes but only an indirect, long-term effect on patient outcomes. The outcome measure was therefore concordance to national CR guidelines regarding the CR program that was prescribed during the NAP. We defined concordance at the level of the patient; it implied prescribing therapy to patients who should be treated and not...
prescribing therapy to patients who should be untreated, according to the guidelines. This was determined for each of the four group-based therapies separately.

**Cluster randomization and allocation**

Randomization of CR clinics was stratified by size (more/less than 30 patients starting CR per month). Per stratum, we generated a randomization scheme with randomly assigned block sizes of either two or four CR clinics using dedicated software. This scheme was concealed to those enrolling and assigning CR clinics [7]. Due to the character of the intervention, it was not possible to blind participants or those involved in providing the intervention.

**Statistical analysis**

For each of the four CR therapies (education, lifestyle modification, exercise training, and relaxation training) we performed a separate mixed-effect logistic regression analysis [10, 23] to assess the effect of the intervention on concordance with guideline recommendations. To this end we included covariates study arm, time, and the interaction between study arm × time. We focused on the interaction term to assess the difference in change over the study period between the two arms—that is, the effect of the intervention—because we expected concordance to improve gradually. We used random effects to model the variation in baseline concordance between clinics (random intercept for each clinic) and the variation in change in concordance over time (random slope for time). To adjust for differences in case mix between the study arms, we included in our analysis three patient level variables (age, sex, and indication for CR) and two clinic level variables (weekly volume of new patients, and whether the clinic is a specialized rehabilitation center or part of a university or teaching hospital) as covariates.

Patients who were seen in the last month of a clinic’s study period were excluded from the analysis because their prescription data was often not yet complete. We also excluded patients for whom the indication for CR was missing. Furthermore, for each of the four analyses of guideline concordance on specific CR therapies we excluded patients for whom it could not be determined whether the prescription of that therapy was concordant with the guideline (either because the guideline’s recommendation or the actual prescription could not be determined).

![Figure 1. Flow diagram of CR clinics through the trial.](image-url)
Results

Participants

Of the 22 eligible CR clinics 18 clinics accepted our invitation to participate in the trial, of which twelve clinics were assigned to intervention arm A (receiving multifaceted A&F intervention on psychosocial therapies), and six to arm B (receiving intervention on physical therapies) (see Figure 1). CR clinics were enrolled in the study between July 2012 and December 2013. On average, the time between randomization and the first educational visit was 3.5 months (standard deviation [SD] 0.7). The average time between subsequent visits was 4.0 months (SD 1.4). Table 1 shows the baseline characteristics of clinics and patients. During the study period a total of 14,847 patients were seen for a NAP in the participating CR clinics. After exclusions in the overall database 11,932 patients were included for the analyses on concordance per CR therapy. The analyses were performed on data from 10,730 (education), 10,774 (lifestyle modification), 10,953 (exercise training), and 8,804 (relaxation training) patients.

Implementation of the intervention

Table 2 shows detailed information on how, and to what extent, the main components of the A&F intervention were implemented in the participating clinics. Due to limited availability of the QI team, one clinic in arm A completed only three A&F iterations during the study period instead of four. There were no differences in QI team size, number of indicators selected as QI goal and number of actions per goal in the QI plan, attendance to the visits, and mean study period between the two study groups. Attendance to the visits remained the same during the study period. The mean number of selected QI goals in each QI plan decreased from 8.0 (SD 2.4) during the initial A&F iteration to 5.0 (SD 3.2) in the final iteration. QI teams in both groups reportedly resolved 1.8 of these QI goals per A&F iteration, on average.

Table 1. Baseline characteristics of clinics (N=18) and patients (N=11,932) per study arm; values are numbers (%), unless indicated otherwise.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Arm A (A&amp;F on psychosocial domain)</th>
<th>Arm B (A&amp;F on physical domain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number participating</td>
<td>12 (66.6)</td>
<td>6 (33.3)</td>
</tr>
<tr>
<td>Median (min-max) number of patients per year</td>
<td>431 (183–1,156)</td>
<td>370 (256– 988)</td>
</tr>
<tr>
<td>Stratum 'large' (&gt;30 patients monthly starting CR)</td>
<td>6 (50.0)</td>
<td>3 (50.0)</td>
</tr>
<tr>
<td>Use of EPR1</td>
<td>9 (75.0)</td>
<td>1 (16.7)</td>
</tr>
<tr>
<td>CR outpatient clinic type:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-teaching hospital</td>
<td>7 (58.3)</td>
<td>3 (50.0)</td>
</tr>
<tr>
<td>Teaching hospital</td>
<td>2 (16.7)</td>
<td>3 (50.0)</td>
</tr>
<tr>
<td>University hospital or specialized rehabilitation</td>
<td>3 (25.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Patients</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number included in analyses</td>
<td>7,692</td>
<td>4,240</td>
</tr>
<tr>
<td>Mean (SD) age in years</td>
<td>64.9 (11.4)</td>
<td>65.8 (11.8)</td>
</tr>
<tr>
<td>Male gender</td>
<td>5,533 (71.9)</td>
<td>3,027 (71.4)</td>
</tr>
<tr>
<td>Indications for CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACS with revascularization</td>
<td>4,446 (57.8)</td>
<td>2,489 (58.7)</td>
</tr>
<tr>
<td>ACS without revascularization</td>
<td>440 (5.7)</td>
<td>386 (9.1)</td>
</tr>
<tr>
<td>Elective CABG or valvular surgery</td>
<td>1,262 (16.4)</td>
<td>598 (14.1)</td>
</tr>
<tr>
<td>Elective PCI</td>
<td>517 (6.7)</td>
<td>321 (7.6)</td>
</tr>
<tr>
<td>Other elective interventions</td>
<td>341 (4.4)</td>
<td>113 (2.7)</td>
</tr>
<tr>
<td>CHF or stable AP, no intervention</td>
<td>252 (3.3)</td>
<td>179 (4.2)</td>
</tr>
<tr>
<td>Other diagnosis, no intervention</td>
<td>434 (5.6)</td>
<td>154 (3.6)</td>
</tr>
</tbody>
</table>

Abbreviations: A&F= audit and feedback, ACS= acute coronary syndrome, AP= angina pectoris, CABG= coronary artery bypass graft surgery, CHF= chronic heart failure, CR= cardiac rehabilitation, EPR= electronic patient record, PCI= percutaneous coronary intervention, QI= quality improvement, SD= standard deviation.
Effect of the intervention

Table 3 compares crude concordance rates between baseline (first three months) and follow-up (remaining time in the study period) for each of the four therapies. Despite random allocation of the participating clinics into two study arms, Chi-squared testing showed significant differences in baseline concordance for lifestyle modification (p<0.001), exercise (p=0.004), and relaxation training (p<0.001) between the two study groups. Table 3 also presents the results of the mixed-effect logistic regression analyses, which compare the trend in concordance over time between intervention and control groups while adjusting for patient age, sex, and indication for CR and adjusting for clinic type and weekly patient volume. No significant differences were found for any of the four therapies. For three of the four therapies (education, lifestyle modification, and exercise training) there were few missing data (around 10%) with respect to recommended and prescribed care, but for relaxation training we found missing data in 26.2% of cases. This was due to six clinics having substantially lower data quality for the relaxation therapy. A sensitivity analysis in which we excluded these clinics from our dataset did not yield different results.

Discussion

Summary of findings

Our multifaceted A&F intervention did not increase concordance of prescribed CR therapies with guideline recommendations. There appeared to be a high variation in baseline performance and data quality between participation CR clinics. Especially for the relaxation training we had a high percentage of missing data on guideline concordance. Although our intervention facilitated active engagement of local multidisciplinary QI teams in setting their own performance improvement goals, the teams often did not succeed in completing the actions that were needed to achieve those goals.

Strengths and weaknesses of this study

The main strength of our study is that we designed our multifaceted A&F intervention based on both existing knowledge from the literature on effective characteristics of A&F interventions [1-5] and an extensive analysis of potential barriers to further increase guideline concordance in the field of CR [11]. The use of CARDSS Online combined with outreach visits actively involved local QI teams, including managers and cardiologists, in the improvement process. By the development and regularly update of a QI plan with concrete, self-formulated goals, we focused on the decision-making processes at the organizational level to create the necessary conditions for improving guideline concordance. Also, as participating clinics were already working with an EPR with CDS functionality, they did not need to change their workflow to participate in the study and collect data. This pragmatic aspect of our study design may have optimized CR clinics’ willingness to participate and minimized the loss to follow-up. Although this resulted in a relatively large sample of CR clinics in which all different CR clinic types were represented, there were large differences in baseline performance between participants.

A limitation of our study is that only CR clinics that used an EPR with CDS that facilitates registration of our study dataset were eligible to participate. These clinics needed to be willing to share their data for research and to allocate resources to establish a QI team. This potentially resulted in a volunteer bias, as eligible CR clinics were less likely to be understaffed and more likely to have information technology to facilitate routine collection of CR data. The generalizability of our results may thus be limited to clinics that are motivated and equipped to systematically monitor and improve the quality of care they deliver. Second, the intervention allowed QI teams to formulate any improvement actions, even if those were not specifically targeted at improving concordance to a specific guideline recommendation. Although this may have optimized the engagement of the team in the improvement process and the commitment to goal attainment, it undermined the connection between the intervention and our primary outcome measure. This link was further diluted because the set of quality indicators chosen by the expert panel did not include guideline concordance for prescribed CR therapies as a separate indicator. However, we did include concordance statistics on each of the four therapies in our feedback reports. In addition, clinics might have started to improve both CR domains and not just the domain covered in their study arm because the intervention has raised their overall awareness for QI. Last, sometimes there are discrepancies between prescriptions and therapies that are actually received by patients caused by e.g. quality of the content of therapies or patient motivation. The effect of our intervention on received rather than prescribed CR therapies might be different.
Table 2: Implementation of the multifaceted A&F intervention in daily practice per study arm; values are mean (SD) unless indicated otherwise.

<table>
<thead>
<tr>
<th>Implementation of the A&amp;F intervention</th>
<th>Arm A (A&amp;F on psychosocial domain)</th>
<th>Arm B (A&amp;F on physical domain)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Range</td>
<td>Range</td>
</tr>
<tr>
<td>QI teams</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of study period per clinic in months</td>
<td>19.8  (6.0)</td>
<td>12 – 30</td>
</tr>
<tr>
<td>Number of A&amp;F iterations</td>
<td>4.6   (1.0)</td>
<td>3 – 6</td>
</tr>
<tr>
<td>Size of local multidisciplinary QI team</td>
<td>7.5   (2.8)</td>
<td>3 – 13</td>
</tr>
<tr>
<td>Number of QI team members attending outreach visits</td>
<td>5.4   (1.9)</td>
<td>1 – 11</td>
</tr>
<tr>
<td>Number (%) of QI teams receiving first telephone follow-up</td>
<td>5 (41.7)</td>
<td>n.a.</td>
</tr>
<tr>
<td>Number (%) of QI teams receiving second telephone follow-up</td>
<td>3 (25.0)</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

QI action planning

|                                        |       |       |       |       |
| Number of indicators selected as QI goal in QI plan | 6.9  (3.1) | 1 – 14 | 6.3  (2.5) | 0 – 10 |
| Mean number of actions per QI goal in QI plan | 1.9   (0.5) | 1.0 – 3.3 | 1.6  (0.4) | 1.0 – 2.6 |
| Number of achieved QI goals per follow-up A&F iteration | 1.7   (1.5) | 0 – 5 | 1.9  (1.5) | 0 – 6 |
| Number of unachieved QI goals in final A&F iteration | 5.9   (3.5) | 1 – 13 | 3.5  (2.2) | 0 – 7 |

Abbreviations: A&F= audit and feedback, CDS= computerized decision support, CR= cardiac rehabilitation, n.a.= not applicable, QI= quality improvement, SD= standard deviation.

Table 3: Concordance rates and difference in concordance between study arms for the four prescribed CR therapies (N=12,111).

<table>
<thead>
<tr>
<th>CR therapies</th>
<th>Crude concordance at baseline a) Intervention</th>
<th>Crude concordance at follow-up b) Intervention</th>
<th>Adjusted odds ratio (95% CI) c)</th>
<th>N (Clinics)</th>
<th>Missing values d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychosocial domain (A&amp;F intervention for arm A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>87.5% (1,141/1,612)</td>
<td>81.2% (1,411/1,612)</td>
<td>90.4% (5,045/5,584)</td>
<td>71.3% (2,191/3,072)</td>
<td>1.28 (0.65 to 2.54)</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>63.0% (1,023/1,624)</td>
<td>34.3% (155/452)</td>
<td>63.2% (3,548/5,610)</td>
<td>25.9% (800/3,088)</td>
<td>0.75 (0.14 to 4.03)</td>
</tr>
<tr>
<td>Physical domain (A&amp;F intervention for arm B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exercise</td>
<td>82.6% (399/483)</td>
<td>89.9% (1,460/1,625)</td>
<td>83.6% (2,667/3,191)</td>
<td>95.1% (5,378/5,654)</td>
<td>0.58 (0.24 to 1.38)</td>
</tr>
<tr>
<td>Relaxation</td>
<td>38.8% (124/320)</td>
<td>72.6% (976/1,345)</td>
<td>51.6% (1,211/2,348)</td>
<td>82.5% (3,952/4,791)</td>
<td>0.4 (0.06 to 2.79)</td>
</tr>
</tbody>
</table>

Abbreviations: A&F= audit and feedback, CDS= computerized decision support, CI= confidence interval, CR= cardiac rehabilitation.

a) Observed concordance during first 3 months of study period.
b) Observed concordance after baseline period until end of study.
c) Odds ratio of improvement in guideline concordance after receiving the A&F intervention for 1 year versus no intervention; adjusted for patients’ age, gender, indication for CR, and clinics’ type and size.
d) Patients for whom the CDS could not provide advice caused by missing data and/or it was not recorded whether the therapy was included in the patients’ CR program.
Strengths and weaknesses in relation to other studies

Our multifaceted A&F intervention included the development and revision (up to five times) of a QI plan based on indicator-based performance in quarterly feedback reports. Such iterative cycles and repeated use of data over time are generally considered key features to improve health care processes. However, a recent systematic review of studies employing the plan-do-study-act method showed that less than 20% of such studies use iterative cycles of change, and only 14% of them repeatedly use data over time [24].

Furthermore, the combination of A&F with both web-based guidance through the process of systematically developing a QI plan and outreach visits to encourage the local QI team to regularly monitor the feedback and update their plan, stimulated engagement of the QI team. Although this is a known characteristic of effective A&F interventions [1-5], other studies struggle with active engagement of health care professionals in goal setting and action planning to improve their performance [25-27]. Ivers et al [27] performed a qualitative study to understand the usefulness of A&F among family physicians and examined barriers to using it to improve quality of care. Their main findings address some general concerns during implementation of A&F interventions to improve professional performance. Participants reported that the feedback increased their awareness of gaps between ideal and actual performance. This resulted mainly in efforts to “try harder” patient by patient. Key barriers to acting upon feedback in a systematic manner included a perceived discordance between population-level quality targets and patient-centered care (“It [A&F] talks about whole populations as opposed to the one individual and I think my approach to this job is the one person at a time”), as well as competing priorities at both the patient and organizational levels (“How much time do you want your doctor devoting to that [A&F], because the more … the less time I am [devoting] to the patient”). A qualitative analysis which is currently underway should point out if similar barriers were present during the implementation of our multifaceted A&F intervention.

Meaning and implications of findings

Further analyses should point out whether participating clinics were able to improve their performance on individual quality indicators, and whether this was related to the selection of these quality indicators in QI plans and to achieving self-formulated improvement goals. If there was indeed improvement on individual indicators, the failure to achieve progress in concordance of prescribed CR therapies with guideline recommendations is probably due to a poor link between these indicators and guideline concordance of therapeutic prescriptions. If there is no improvement on individual indicators, then our A&F intervention has simply failed to stimulate clinicians to work on QI actions outside their daily routine. The large number of unattained QI goals (Table 2) points in this direction.

According to Ancker et al [28] the evaluation of health information systems, like our web-based A&F intervention, often show mixed results. This may be in part attributable to the evaluation frameworks used. They developed a model for evaluation, named the Triangle Model, in which they emphasize the sociotechnical view that organization, technology, and users influence and change each other during implementation processes. The lack of success of our web-based A&F intervention might not have only depended on the technology used but also on the organizations and professionals involved. Similarly, the Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) framework [29] emphasizes that a thorough needs analysis should be performed to determine organizational readiness before initiating change. This might uncover underlying issues within the institution (e.g. equipment problems or staffing shortages) which first should be resolved to make the QI effort succeed [29]. Also the Systems Engineering Initiative for Patient Safety (SEIPS) model [30] presents a broad approach with a focus on system design and its impact on processes and outcomes. This model describes the structure of a health care organization as a work system with five components (person, tasks, tools and technologies, physical environment, organizational conditions) who interact with each other and affect both work (e.g. maintenance and supply chain management) and clinical care processes. Both processes in turn influence the patient, employee, and organizational outcomes of care [30]. Capturing more detailed predictor variables about the technology, users, and the surrounding context might have increased the ability to interpret our findings of process variables (e.g. organization – professional processes such as culture and workflow [28]) during the evaluation of our intervention.

Future work

Although A&F interventions are increasingly used to aid health care professionals in improving their performance, they might not qualify as the best basis for improving concordance of prescribed therapies with guideline recommendations. Our future work will include results on the intervention’s effect on concordance of the received CR therapies with guideline recommendations, as well as results on team performance (the intervention’s effect on all quality indicators); and results of a qualitative process evaluation. During this evaluation we use the concept
mapping methodology (including focus group sessions) to explore experiences from participating CR clinics with
the intervention to gain insight into barriers and facilitators of the implementation.

Conclusion

A web-based A&F intervention with outreach visits did not increase concordance of prescribed CR therapies with
guideline recommendations in a pragmatic evaluation using EPRs for data collection. There appeared to be a high
variation in baseline performance and in data quality among participating CR clinics. Although QI teams in the
clinics formulated QI goals and associated actions at the start of each quarterly A&F iteration, most goals were not
attained. We recommend to align data registration in participating clinics before starting an A&F intervention that
uses EPRs for data collection. Future analyses should show whether our intervention did improve the overall CR
team performance measured by change in quality indicators results, complemented with qualitative information on
factors which influenced the implementation of the A&F intervention.

References

1. Ivers N, Jamtvedt G, Flottorp S, Young JM, Ogdgaard-Jensen J, French SD, et al. Audit and feedback: effects
behaviour change interventions: the example of audit and feedback. Social science & medicine (1982).
2010;70(10):1618-25.
5. van der Veer SN, de Keizer NF, Ravelli AC, Tenkink S, Jager KJ. Improving quality of care. A systematic
review on how medical registries provide information feedback to health care providers. Int J Med Inform.
6. Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW. Improving quality improvement
Evaluating the effect of a web-based quality improvement system with feedback and outreach visits on
guideline concordance in the field of cardiac rehabilitation: rationale and study protocol. Implementation
8. van Engen-Verheul MM, van der Veer SN, de Keizer NF, Tjon Sjoe Sjoe W, van der Zwan EP, Peek N. A
Web-based System to Facilitate Local, Systematic Quality Improvement by Multidisciplinary Care Teams:
computerised decision support on decision making of multidisciplinary teams: cluster randomised trial in
cardiac rehabilitation. BMJ. 2009;338:b1440.
computerised decision support on barriers to guideline implementation: a qualitative study in outpatient cardiac
prevention in the clinical management of patients with cardiovascular diseases. Core components, standards
and outcome measures for referral and delivery: A Policy Statement from the Cardiac Rehabilitation Section of
the European Association for Cardiovascular Prevention & Rehabilitation. Endorsed by the Committee for
Practice Guidelines of the European Society of Cardiology. European journal of preventive cardiology.
measures on cardiac rehabilitation for referral to and delivery of cardiac rehabilitation/secondary prevention
services endorsed by the American College of Chest Physicians, American College of Sports Medicine,
American Physical Therapy Association, Canadian Association of Cardiac Rehabilitation, European
Association for Cardiovascular Prevention and Rehabilitation, Inter-American Heart Foundation, National


Biological Model Development as an Opportunity to Provide Content Auditing for the Foundational Model of Anatomy Ontology

Lucy L. Wang, MS¹, Eli Grunblatt², Hyunggu Jung, MS¹, Ira J. Kalet, PhD¹,³, Mark E. Whipple, MD, MS¹,⁴

¹Department of Biomedical Informatics and Medical Education
²Medical Scientist Training Program
³Department of Radiation Oncology
⁴Department of Otolaryngology-Head and Neck Surgery
University of Washington, Seattle, WA

Abstract
Constructing a biological model using an established ontology provides a unique opportunity to perform content auditing on the ontology. We built a Markov chain model to study tumor metastasis in the regional lymphatics of patients with head and neck squamous cell carcinoma (HNSCC). The model attempts to determine regions with high likelihood for metastasis, which guides surgeons and radiation oncologists in selecting the boundaries of treatment. To achieve consistent anatomical relationships, the nodes in our model are populated using lymphatic objects extracted from the Foundational Model of Anatomy (FMA) ontology.

During this process, we discovered several classes of inconsistencies in the lymphatic representations within the FMA. We were able to use this model building opportunity to audit the entities and connections in this region of interest (ROI). We found five subclasses of errors that are computationally detectable and resolvable, one subclass of errors that is computationally detectable but unresolvable, requiring the assistance of a content expert, and also errors of content, which cannot be detected through computational means. Mathematical descriptions of detectable errors along with expert review were used to discover inconsistencies and suggest concepts for addition and removal. Out of 106 organ and organ parts in the ROI, 8 unique entities were affected, leading to the suggestion of 30 concepts for addition and 4 for removal. Out of 27 lymphatic chain instances, 23 were found to have errors, with a total of 32 concepts suggested for addition and 15 concepts for removal. These content corrections are necessary for the accurate functioning of the FMA and provide benefits for future research and educational uses.

Introduction
The Foundational Model of Anatomy (FMA) ontology¹ seeks to represent knowledge within the domain of human anatomy in a symbolic way. It attempts to formalize all parts and relationships in the body, and intends to be a “resource for developing the anatomy content of applications that target specific user groups,” which includes biomedical research applications.¹ A human biological model may focus on the anatomy and physiology of a particular organ system or body part, and may therefore only concern itself with a subset of elements from the FMA. Isolating and using these specific concepts from the FMA allows for better linkage and recycling of models, as the terminology of the FMA can be used to bridge gaps between the authors of these models. More importantly, perhaps, and the focus of this paper, model building also helps to enhance the content of the FMA itself. Certain weaknesses in ontological structure as well as internal inconsistencies can be brought to light through the demands of practical application.

In this paper, we focus on a specific use case as an example: modeling tumor dissemination in the lymphatics of the head and neck. HNSCC is a major form of cancer that arises in the mucosa of the upper aerodigestive tract. This particular form of cancer is known for a high propensity towards metastasis, with the spread of tumor cells occurring nearly universally through the lymphatic system.²,³ For HNSCC, standard treatment involves surgical resection and/or irradiation of part of the neck to remove the mass and lymph nodes in which metastasis is suspected. Because both surgical removal and over-irradiation of surrounding tissue have negative consequences to the patient, it is in the clinician’s interest to minimize collateral damage. Ideally, only lymphatic regions with high probabilities of microscopic disease would be treated. Determining the locations within the head and neck with high likelihood for metastasis is critical in guiding surgeons and radiation oncologists in selecting the boundaries of treatment.⁴–⁶ In this region of the body, the metastatic trends generally follow known, but not precisely defined paths.⁷,⁸ Markov chain models have shown promise in quantifying the probabilities of metastasis to particular nodal regions⁹; we attempt to build such a model using the lymphatic topology represented in the FMA.
The anatomical underpinning of the model is retrieved from the FMA ontology. This allows us to establish consistency between our model and existing knowledge about lymphatic anatomy. Nodes in our Markov chain model are populated using terms and relationships derived from the relevant portions of the FMA. In most cases, the ontology provides a sufficient representation of the regional lymphatics. However, we were able to determine several areas of insufficiencies. These include failures to distinguish between instances and superclasses of lymphatic objects, missing concepts, and to a much lesser degree, incorrect concepts or connections. For many of these issues, we were able to fill in the gaps of knowledge using existing information in the FMA (i.e., by relocating mis-located information, or inferring information from existing relationships). In other cases, we contended with issues in the fundamental organization of the FMA, and a more significant discourse was necessary.

Our strategies for error detection and correction can be used to improve other parts of the FMA ontology. Some of our methods are outlined below; it is our hope that these techniques can be used to assess other parts of the FMA and bring about changes which will ultimately improve its logical underpinnings, consistency and usefulness.

Methods
We evaluate errors in the FMA ontology that can be categorized into the following classes: (1) computationally detectable and resolvable, (2) computationally detectable and unresolvable, or (3) undetectable. All errors can be addressed through the addition or removal of select concepts. Errors of class 1 and 2 can be described in logical terms and algorithmically discovered. Of these, fixes for class 1 errors can be constructed from pre-existing information in the FMA ontology. Fixes for class 2 errors cannot be automatically generated, and these errors must be addressed by content experts. Lastly, errors of class 3 can neither be detected nor fixed using computational methods. These are errors of content, where the ontology disagrees with established literature. These errors can currently only be detected through systematic validation by an anatomical expert. In the construction of our model, we were able to discover and determine fixes for all three classes of errors described above. Computational methods for automatic error detection were attempted in all cases. Once all possible detectable errors were found and fixed, an anatomical expert assessed and completed the auditing of the edited ontology.

We begin by describing the role of the FMA in the construction of our Markov chain model. Our ROI includes parts of the head and neck which constitute the mucosa of the upper aerodigestive tract, the lymphatic objects that drain these parts, and their lymphatic connections to venous circulation. We follow by describing the classes of errors that we found in this region, and how we infer concepts for addition and removal for each error class.

All anatomical parts, their names and properties were retrieved from the latest release of the FMA in the Web Ontology Language (OWL), version 3.2.1 (University of Washington Structural Informatics Group (UW-SIG)). SPARQL (SPARQL Protocol and RDF Query Language) queries were written and stored in the Query Integrator (UW-SIG) and accessed through an interface written in Clozure Common Lisp version 1.7 (Clozure Associates, Brookline, MA). All graphs of the lymphatic network were generated using Graphviz version 2.36 (AT&T Labs, Austin, TX).

Using the FMA to construct our model
The process of cancer metastasis can be represented using Markov chains, a type of memoryless stochastic model useful for describing the probability of events based on previous known states. For our Markov chain model of metastasis, each node captures both the anatomical location and the severity of disease. In cases of HNSCC, metastasis occurs along known lymphatic drainage paths. We can therefore model locations as lymphatic concepts from the FMA and severity using the cancer T-stage grading system.

Lymphatic flow can be modeled as a unidirectional process. The lymph fluid is assumed to flow only in the direction of the two great lymphatic vessels: the thoracic duct and the right lymphatic duct, before entering venous circulation. At any point in time, all nodes downstream of the primary tumor site have some probability of hosting metastatic disease. Our model therefore must consist of all lymphatic structures reachable by a particular primary tumor. Because there are known connections between identifiable lymphatic regions, we can construct a partial map of lymphatic drainage in the head and neck. We do this by extracting the relevant anatomical concepts and connections from the FMA.

The lymphatic network in the FMA is organized under the terms trees, trunks, branches, tributaries, chains and nodes. The right and left lymphatic trees are considered organs, and the subbranches and subtrees that form them are considered organ regional parts. Within these lymphatic trees, the trunk identifies the main stem of the tree. Further bifurcation of the stem yields branches and tributaries. These further consist of chains, whose members are nodes. Although this mode of organization may not be analogous to all accepted conceptualizations of the lymphatic system,
Figure 1. Values under the lymphatic_drainage property of the FMA object Soft_palate are queried and used to generate a full map of lymphatic flow downstream from the origin site. All unlabeled arrows between lymphatic objects represent the efferent_to relationship.

it does provide an appropriate number of subdivisions which can be used in modeling. The states in our Markov chain model refer to lymphatic chains, the smallest subdivision that is well-represented in the FMA. For the purposes of this article, specific entities and relationships from the FMA will be referred to in italics, with underscores in place of spaces.

Each tumor has an identifiable origin site, which corresponds to an entity in the FMA (e.g. Tongue, Soft_palate, Floor_of_mouth etc). Organ entities have a property lymphatic_drainage whose values are the lymphatic objects that drains the site. A SPARQL query can be written to extract this information, returning all lymphatic objects which directly drain an organ or organ part. For example, if a tumor is located in the tongue, we expect a query to retrieve the lymphatic objects \{Right_jugulodigastric_lymphatic_chain, Left_jugulodigastric_lymphatic_chain, Submental_lymphatic_chain, Basal_lingual_lymphatic_tree, Central_lingual_lymphatic_tree, Right_marginal_lingual_lymphatic_tree, Left_marginal_lingual_lymphatic_tree\}. The number of results can then be reduced by increasing the detail of the query, i.e., by querying on a regional part of the tongue, such as the Apex_of_tongue or Body_of_tongue. The FMA should be able to accommodate the most specific and anatomically relevant organ part, and return only those results appropriate for that entity. As you will see shortly, the annotation on the tongue is incomplete, and some modifications must be made to achieve these expected results.

Once we determine the direct lymphatic drainage of the tumor origin, we set out to define all possible paths of drainage. The FMA defines a relationship efferent_to for all lymphatic objects. This property points to lymphatic chains which are downstream of the chain of interest, i.e., closer to the great lymphatic vessels. Chains with multiple values under this property represent branching points in the lymphatic network. A full drainage map can be constructed by querying recursively over all efferent lymphatic chains until each one terminates in one of the two great lymphatic vessels. By combining the lymphatic objects at the tumor origin and the chains defined by the efferent_to relationship, we form a map defining all possible routes of metastasis for a particular tumor through the lymphatic network. The lymphatic objects in the final map can then be used to define nodes in the Markov chain model. Figure 1 shows the lymphatic drainage map constructed using entity Soft_palate as the seed tumor origin.

Inconsistencies identified in the FMA representation of the lymphatics
The FMA provided in large part a description of the lymphatic network which was accurate and detailed enough for our needs, such as the output presented in Figure 1. However, in many cases, the methods described above did not generate
drainage paths which correctly represented lymphatic anatomy. Inconsistencies arose in the connections between lymphatic chains, as well as in the definition of the lymphatic_drainage property. Five subclasses of class 1 (detectable and resolvable) errors were discovered: (1a) organs whose direct lymphatic drainage incompletely represents the drainage of their regional parts, (1b) organs with lymphatic drainage by lymphatic chain superclasses, (1c) lymphatic objects with efferent connections to inappropriate objects, i.e., superclass objects, (1d) connections between lymphatic chain superclasses, and (1e) erroneous connections between objects in the left and right side lymphatic trees. One subclass of class 2 (detectable and unresolvable) errors was discovered, (2a) lymphatic objects with no efferent connections. Class 3 (undetectable) errors were also found through anatomical expert review following the detection and correction of the previous error subclasses.

We use the following notation to indicate specific types of objects in the ROI. Allow $A$ to be the set of all lymphatic objects in the ROI. $A$ consists of $S$, the set of superclass lymphatic chains, $C$, the set of instances of superclass lymphatic chains, and $G$, the set of great lymphatic vessels, defined as $\{\text{Thoracic duct, Right lymphatic duct}\}$. $S$, $C$ and $G$ are mutually disjoint, and $S \cup C \cup G = A$. The FMA consists of structural entities, material objects in the body, i.e. left/right lung, and abstract entities, i.e. lung. The members of $S$ are abstract entities, whose members’ instances $C$ refer to structural entities in the body. Consequently, each member $s_i$ in $S$ has two instances representing bilaterally placed chains, $s_{iL}$ and $s_{iR}$, the right and left side chains respectively, both of which are members of $C$.

$C$ then consists of $L$, the set of lymphatic chains in the left side body, $R$, the set of lymphatic chains in the right side body, and $M$, the set of lymphatic chain instances that are located medially. $L$, $R$, and $M$ are also mutually disjoint and $L \cup R \cup M = C$. For any lymphatic object $c_i$ in $C$, allow $F(c_i)$ to represent the set of all lymphatic objects with a one-step efferent, to relationship to $c_i$.

Any organ or organ part $N$ consists of recursive regional parts $P = \{p_1, p_2, ..., p_n\}$. Each regional part $p_i$ in $P$ is drained by some set of lymphatic objects $Q_i = \{q_{i1}, q_{i2}, ..., q_{im}\}$. $N$ also has lymphatic drainage objects $D = \{d_1, d_2, ..., d_k\}$, acquired from querying on the lymphatic_drainage property of $N$.

Error constraints are expressed using set theoretic notation in Table 1 for clarity and portability. These are the foundations of queries ultimately used to extract inconsistencies from the FMA. All errors are described in greater detail below.

**Table 1. Descriptions of all error class constraints using set theoretic notation**

<table>
<thead>
<tr>
<th>Error subclass</th>
<th>For any</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>$N$</td>
<td>$\bigcup_{i \in I} Q_i \subseteq D, I = {1, 2, ..., n}$</td>
</tr>
<tr>
<td>1b</td>
<td>$p_i$</td>
<td>$Q_i \cap S = \emptyset$</td>
</tr>
<tr>
<td>1c</td>
<td>$c_i$</td>
<td>$F(c_i) \subset ((G \cup C) \cap {c_i})^c$</td>
</tr>
<tr>
<td>1d</td>
<td>$s_i$</td>
<td>$F(s_i) = \emptyset$</td>
</tr>
<tr>
<td>1e</td>
<td>$r_i, l_i, m_i$</td>
<td>$F(l_i) \subset L \cup M \cup G; \quad F(r_i) \subset R \cup M \cup G; \quad F(m_i) \subset L \cup R \cup M \cup G$</td>
</tr>
<tr>
<td>2a</td>
<td>$c_i$</td>
<td>$F(c_i) \neq \emptyset$</td>
</tr>
</tbody>
</table>

**Subclass 1a** An organ is necessarily drained by all the lymphatic objects which drain the organ’s regional parts. Conversely, however, the lymphatic objects that drain all of an organ’s regional parts may not represent all of the chains that drain the whole organ. For example, a query on the entity Tongue returns the lymphatic drainage object $\{\text{Jugulodigastric_lymphatic_chain}\}$, yet a query on all of the tongue’s regional parts returns $\{\text{Basal lingual_lymphatic_tree}, \text{Marginal lingual_lymphatic_tree}, \text{Central lingual_lymphatic_tree}\}$. The first is not a superset of the second, violating the constraint given above.

**Subclass 1b** No organ or organ part should be drained by superclass lymphatic objects. The entity Tongue is drained by the Jugulodigastric_lymphatic_chain, a superclass object, which violates this constraint.

**Subclass 1c** No lymphatic object should have efferent connections to inappropriate objects such as superclass lymphatic chains. For any lymphatic object $c_i$ in $C$, $F(c_i)$ must contain only appropriate objects such as members of $C$ or
Subclass 1d Superclass chains should not have efferent_to relationships to other objects. The FMA is inconsistent with this distinction and some superclasses are connected erroneously. As an example, the Superior_lateral_deep_cervical_lymphatic_chain has an efferent connection to the Inferior_lateral_deep_cervical_lymphatic_chain. Both of these objects are superclass lymphatics, and the connection is erroneous.

Subclass 1e As above, the right and left sides of this ROI are drained by distinct lymphatic trees. Right side lymphatic chains connect to other right side chains and similarly for the left side. The starting lymphatic object Left_subscapular_axillary_lymphatic_chain, for example, has efferent chains \{Right_central_axillary_lymphatic_chain, Right_apical_axillary_lymphatic_chain\}. The starting object is a member of \(L\), yet both of its efferent objects are members of \(R\), violating our constraint. This example is outside of our ROI, and although no subclass 1e errors were detected within the scope of this paper, we felt it important to include the class definition here for future use.

Subclass 2a All lymphatic objects must have efferent relationships to other lymphatic objects. Examples of lymphatic objects in the FMA with no efferent objects are the left and right submandibular lymphatic chains. Querying the efferent_to property of either returns the empty set. Although we were able to detect these errors, we could not infer corrections from existing concepts in the ontology.

Class 3 Type 3 errors involve incorrect information in the FMA ontology. Certain efferent connections between lymphatic objects may not agree with the prevailing literature.\(^{14-19}\) These content errors must be identified and corrected through expert review.

Using existing information in the FMA to correct inconsistencies

Many of these inconsistencies can be corrected systematically. Subclass 1a errors in which an organ is not correctly labeled with a superset of the lymphatic drainage objects of all of its regional parts can be addressed by adding the missing relationships to the FMA. For each organ in the ROI, we construct two queries: one to find the lymphatics that drain the organ directly, \(D\), and one to find the union of lymphatics which drain all of the organ’s recursive regional parts, \(\bigcup_{i \in I} Q_i\), \(I = \{1, 2, ..., n\}\), which we refer to as \(Q\). If the result of the first query is not a superset of the result of the second query, the lymphatic objects in \(Q \setminus D\) should be added as lymphatic_drainage objects of the organ. This process is then applied to all organ parts and parts of parts to achieve complete annotation.

For subclass 1b errors, we detect all organ or organ parts which are drained by superclass lymphatic objects. For each organ, we use the same query results for \(Q\) given above. Members of set \(Q \cap S\) should be removed as the lymphatic_drainage for the organ. Additionally, for each inappropriate drainage relationship, two new efferent concepts can be inferred. Lymphatic drainage relationships can be inferred between the organ and the left and right side instances of the inappropriately connected superclass object.

Likewise, subclass 1c errors can be detected by querying for efferent connections between lymphatic instances and superclass objects. For each lymphatic object \(c_i\), concepts in the set difference \(F(c_i) \setminus (G \cup C) \cap \{c_i\}\) are marked for removal as efferent_to values of \(c_i\).

We can also use the inappropriate connections in this case to infer correct but missing connections. The entity Central_lingual_lymphatic_tree for example, is a medial lymphatic chain (member of \(M\)). Querying on its efferent_to relationship returns \{Superior_lateral_deep_cervical_lymphatic_chain, Submandibular_lymphatic_chain\}, both of which are superclass lymphatics (members of \(S\)). Although this violates the constraint for subclass 1c errors, the intention of these connections may still be correct. If so, we extrapolate that Central_lingual_lymphatic_tree does in fact connect to both of these efferent chains, but to the right and left instances rather than the superclasses. We can then infer the addition of four efferent connections to:

\{Right_superior_lateral_deep_cervical_lymphatic_chain, Left_superior_lateral_deep_cervical_lymphatic_chain, Right_submandibular_lymphatic_chain, Left_submandibular_lymphatic_chain\}.

Subclass 1d errors can be identified and extraneous connections between lymphatic chain superclasses removed. In cases where the left and right side connections do not exist, the erroneous superclass connection can be used to infer them. After verifying the existence of appropriate connections between the right and left instances and their respective efferent objects, the connections between the superclass lymphatic objects are marked for removal.

Subclass 1e errors can be corrected by changing inappropriate efferent connections to point to the ipsilateral lymphatic chain instance. We identify all lymphatic objects that connect efferently to lymphatic objects on the contralateral side. These connections are marked for deletion. If the connection to the analogous ipsilateral lymph chain does not
already exist, a new efferent connection is added.

Subclass 2a errors cannot be addressed without input from a content expert. On occasion, the FMA harbors incorrect connections to superclasses (subclass 1c or 1d errors) that can be used to infer missing connections. In most cases though, it is impossible to conclude from existing information in the FMA how disjoint lymphatic objects should connect to the lymphatic network. This information must instead be found in the literature and propagated back into the FMA following review. Some examples of disconnected instances in the FMA are the right and left submandibular and the right and left paratrachael lymphatic chains. An existing connection in the FMA between the superclass Submandibular_lymphatic_chain and the Jugulo-omohyoid_lymphatic_chain indicates that this is the potential correct connection for the left and right instances. For the paratrachael chains, however, there is no available superclass connection from which to draw similar conclusions, and an anatomist must be consulted.

Diagrammatic representations of these error subclasses are given in Figure 2. Most errors could be detected and corrected computationally. Those errors that could not (class 3) were detected and corrected by a content expert.

Results

Within the upper aerodigestive tract, we extracted 106 distinct organ and organ regional parts from the FMA. There were 27 distinct lymphatic chain instances (right, left or medial) and 11 distinct superclasses of lymphatic objects in the ROI. Table 2 gives the number of inconsistencies detected for each error subclass, as well as the numbers of concepts suggested for addition and removal. These suggestions were generated first by programmatic rule evaluation and then confirmed with an expert anatomist. Counts in the table labeled with * are based solely on expert review, and represent circumstances in which the FMA differs from anatomical literature. Some suggestion counts may
disagree with the numbers implied in the Methods section, and this is either due to repetition in suggested concepts, or removal of suggestions based on expert review.

**Table 2.** Error counts for each class of error in the ROI.

<table>
<thead>
<tr>
<th>Error class</th>
<th>Affects</th>
<th>Objects affected</th>
<th>Total objects</th>
<th>Concepts to add</th>
<th>Concepts to remove</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>organ (parts)</td>
<td>6</td>
<td>106</td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td>1b</td>
<td>organ (parts)</td>
<td>4</td>
<td>106</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>1c</td>
<td>lymphatic instances</td>
<td>3</td>
<td>27</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1d</td>
<td>lymphatic superflasses</td>
<td>7</td>
<td>11</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>1e</td>
<td>lymphatic instances</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2a</td>
<td>lymphatic instances</td>
<td>12</td>
<td>27</td>
<td>10*</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>lymphatic instances</td>
<td>7*</td>
<td>27</td>
<td>8*</td>
<td>4*</td>
</tr>
<tr>
<td>All</td>
<td>all entities</td>
<td>37</td>
<td>144</td>
<td>62</td>
<td>23</td>
</tr>
</tbody>
</table>

Prior to this round of content auditing, many lymphatics within the ROI were disconnected from the rest of the system (12 out of 27) or were connected erroneously to superclass entities (3 out of 27). Following automated inference and auditing, the number of disconnected entities drops to 4 out of 27, with no entities connected to superclass objects. Further concepts were added or removed by an anatomist, resolving class 3 errors, resulting in a final lymphatic map with no disconnected lymphatic instances. The structure and interconnectivity of the lymphatic ROI is shown in Appendix A. The network is shown in its original form (preceding any auditing) and at the completion of our review. Only right-side and medial lymphatic instances are shown for clarity. Final suggested concepts are in the process of being propagated back into the FMA.

**Discussion**

The FMA has been open to both changes to its underlying structure as well as content auditing. Over the past several years, it has undergone a dramatic change from a Frames-based system to an OWL representation.\(^{20,21}\) This change did not resolve many content issues, but increased the flexibility of the ontology, allowing for easier content auditing. Several attempts have been made in this regard, all of which have led to incremental improvement. In 2009, Gu et al. detected potential incorrect relationships by studying the implicit relationship between the *is-a* statement and structural relations such as *part-of*. An object A which is part of another object B, for example, may not also be an instance of B. Entities found to violate the implicit logic of these relationships were vetted by an expert anatomist and removed or corrected in the FMA.\(^{22}\) In the same year, Kalet et al. specifically suggested the auditing of the FMA’s representation of the lymphatic system.\(^{23}\) A study in 2012 expanded on some of these approaches by specifying and detecting “graph motifs,” small sets of relationships known to be problematic. SPARQL queries were used to fetch fragments of the FMA matching these motifs, and these problem concepts were then validated or changed in the FMA.\(^{24}\) Most recently, Luo et al. used the assumption of structural self-bisimilarity to detect potential anomalies in the FMA. They contend that certain types of relationships are expected to be symmetric on the two sides of the body, and differences in bilateral connectivity may be a strong indicator of error.\(^{25}\)

The techniques used in this paper share similarities to some of those mentioned above. One major difference though, is that the current auditing happened in the context of model creation. The deficiencies found needed to be addressed to allow us to continue to use the FMA as a tool. Addition and removal of suggested concepts were immediately vetted not only by an anatomical expert, but through the rigor of modeling use and testing. The corrected portion of the ontology is therefore less likely to contain significant further errors.

Overall, we believe that the structure, organization and intention of the FMA is a suitable foundation upon which to build our Markov chain model of cancer metastasis. However, the prior state of the FMA was unable to support our modeling needs. In our attempts to use the FMA to populate nodes in our model, we were able to systematically review relevant parts of the anatomical representation, specifically the lymphatic system in the mucosa of the upper aerodigestive tract. The classes of errors identified through this process generated a constraint set which may apply broadly to other parts of the lymphatic tree. It is our hope to extrapolate these auditing techniques to the rest of the lymphatics and other analogous systems such as arterial-venous circulation and the peripheral nerves.

**Limitations**

The methods described above have only been used to audit a small section of the FMA ontology. It is yet unknown whether they are applicable to the rest of the lymphatic system or other analogous systems as represented in the FMA.
For example, in the upper body, relationships between lymphatics generally show bilateral symmetry. In the abdomen and lower body, however, lymphatics drain only into the thoracic duct, and this property no longer holds. Also, the organs are asymmetric within the abdominal cavity, so there are no clear superclass and instance relationships. This more complex topology existing in the rest of the body will require not only a reevaluation of the error classes we have defined in this paper, but will likely introduce opportunities for new error definitions. Likewise, applying such methods to the arterial-venous circulatory system or the peripheral nervous system may yield an analogous but different set of constraints.

This paper focused primarily on classes of errors that can be defined and identified computationally. Some errors, although detectable, cannot be automatically corrected. These content errors benefit from automated content auditing by being discovered, but we have no way of proceeding without review by a content expert. We touched on this briefly when discussing subclass 2a errors. For these errors, we cannot infer the correct efferent connections to a disconnected lymphatic object; deductive reasoning will not suffice. We attempt to infer some possible solutions for subclass 2a errors using other pieces of information that already exist in the FMA, such as erroneous connections between superclasses. However, these inferences may not hold in all cases. Often, we cannot infer any new knowledge at all. This limitation is foreseen and expected. After all, we attempt only to detect connections which are potentially erroneous, and not to automatically generate novel content.

Likewise, errors of content, which are not logical in nature and cannot be expressed in mathematical formulae, are extremely difficult to find in an ontology the size of the FMA. Although we were able to discover and fix some of these content errors by hand, the process was laborious and would not scale well to larger subsets of the FMA. Our techniques may therefore face additional challenges when applied broadly to the FMA ontology.

Additionally, we were unable to audit certain other classes of content errors within the scope of this paper. For example, an organ may be fully annotated while none of its parts are. This is of concern when information is missing at our specified resolution, e.g., when we want to know the lymphatic drainage of an organ part, but there is no drainage information at that level in the FMA. We have yet to determine a cheap or systematic way to discover these types of ontological errors.

Conclusions

Many models of biological processes are fundamentally rooted in anatomy. In order to be “anatomically correct,” a model should satisfy the relational and spatial constraints imposed by human anatomy. For example, the human heart has a set of properties, such as having two atria, two ventricles and four valves, and would not be considered a proper heart if these statements were untrue, or if the relative locations of these objects were incorrect. Likewise, in building a functional model of a heart, the modeler would also want to include all of the above constituent parts and their relationships. Otherwise, there is a high probability that the model is inaccurate. Although this is an overly simple example, one could conceive of assessing the accuracy of a model by judging the correctness of its represented anatomy. A mismatch between a model and the known anatomical worldview would not necessarily invalidate the model, but would indicate potential points of improvement.

An ideal anatomical ontology would be able to fulfill these needs. The FMA, in its conception, attempts to represent human anatomy correctly and faithfully. The current state of the FMA has certain insufficiencies that increase barriers to its use, but many of these issues can be addressed through systematic content auditing. The best way to reach the correct anatomical description perhaps, is to selectively use, validate and correct subsets of the FMA ontology. Making edits in such a piecemeal way may not seem like the most efficient course of action, but what it lacks in breadth it makes up for in practicality.

Acknowledgements

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R21-LM012075. This study was supported in part by the National Library of Medicine (NLM) Training Grant T15LM007442. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Many thanks to Onard Mejino and the UW Structural Informatics Group for providing insight into the content and organization of the FMA.

References


10 FMA in OWL [Internet]. University of Washington Structural Informatics Group;.


2119
Appendix A

Connected graphs of the right side lymphatic network: the original extracted from FMA (top), and the final audited version (bottom).
Automated Extraction of Substance Use Information from Clinical Texts

Yan Wang, MS1, Elizabeth S. Chen, PhD4,5, Serguei Pakhomov, PhD1,2
Elliot Arsoniadis, MD1,3, Elizabeth W. Carter, MS4, Elizabeth Lindemann1,
Indra Neil Sarkar, PhD, MLIS4,6, Genevieve B. Melton, MD, PhD1,3

1Institute for Health Informatics, 2College of Pharmacy, and 3Department of Surgery,
University of Minnesota, Minneapolis, MN; 4Center for Clinical & Translational Science,
Department of Medicine, 5Department of Microbiology & Molecular Genetics,
University of Vermont, Burlington, VT

Abstract
Within clinical discourse, social history (SH) includes important information about substance use (alcohol, drug, and nicotine use) as key risk factors for disease, disability, and mortality. In this study, we developed and evaluated a natural language processing (NLP) system for automated detection of substance use statements and extraction of substance use attributes (e.g., temporal and status) based on Stanford Typed Dependencies. The developed NLP system leveraged linguistic resources and domain knowledge from a multi-site social history study, Propbank and the MIPACQ corpus. The system attained F-scores of 89.8, 84.6 and 89.4 respectively for alcohol, drug, and nicotine use statement detection, as well as average F-scores of 82.1, 90.3, 80.8, 88.7, 96.6, and 74.5 respectively for extraction of attributes. Our results suggest that NLP systems can achieve good performance when augmented with linguistic resources and domain knowledge when applied to a wide breadth of substance use free text clinical notes.

Introduction
Social history (SH) factors including alcohol, drug, and nicotine use (collectively referred to as “substance use”) are increasingly recognized as risk factors for preventable disease, disability, and mortality. A number of studies have been published describing the linkage between social risk factors and their associated morbidity or mortality(1-4). For example, nicotine abuse continues to be the leading preventable cause of morbidity and mortality in the United States(5), and the severity of substance use disorders is strongly associated with the magnitude of comorbidity(3). In 2006, the Institute of Medicine (IOM) report on “Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate” described the need for improving existing datasets, developing new data sources, and establishing strategies and models for incorporating behavioral and environmental factors and their interactions(6). In realizing this important link, the last decade has had increasing attention on substance use in its many forms with respect to health and disease(6-11).

The availability of electronic documents within electronic health record (EHR) systems provides an opportunity for clinical researchers to access a wide range of information about an individual’s social environment and use of this information for secondary purposes (e.g., disease surveillance or evidence-based medicine) as well as primary uses for patient care (e.g., decision support). However, large amounts of detailed substance use information in EHR systems are stored predominantly in free-text rather than structured format(12, 13), underscoring the need for automated information extraction or other natural language processing (NLP) techniques specific for identifying social history information such as substance use.

In this study, we sought to develop an NLP system for detecting three main sub-categories of substance use (alcohol, drug and nicotine use) statements within free-text clinical notes and extracting related information (e.g., temporal, amount and type) within these substance use statements.

Background

Substance use information in clinical documents and structured modules in the EHR

This research group has previously performed a multi-institutional study on social history information in EHR system clinical notes(14). In this initial study, social history information contained in clinical notes from different sources (MTSamples(15), University of Vermont Medical Center [UVMMC; formerly Fletcher Allen Health Care], and University of Minnesota affiliated Fairview Health Services [FHS]) was analyzed and the adequacy of HL7 CDA-based models(16) and openEHR(17) archetypes for representing social history information within relevant statements across the three sources was studied. Table 1 shows an example of the alcohol use statement type identified in the study along with categories of information within statements for this type and a common set of data
elements and values. Table 2 illustrates the representation of information within a social history statement for nicotine use.

Among modules for EHR substance use, two follow-up studies showed that free-text comments with alcohol and nicotine use were often added due to the limited ability of structured parts of the EHR to describe some elements (e.g., amount and frequency) for both alcohol and nicotine usage (12, 13).

Table 1. Elements and values for alcohol use statement type.

<table>
<thead>
<tr>
<th>SH statement type</th>
<th>Elements</th>
<th>Example value or pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol use</td>
<td>Status</td>
<td>current, past, nondrinker</td>
</tr>
<tr>
<td></td>
<td>Temporal</td>
<td>[in/since/until] &lt;date&gt;</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td>consume, use, drink</td>
</tr>
<tr>
<td></td>
<td>Type</td>
<td>wine, alcohol, beer</td>
</tr>
<tr>
<td></td>
<td>Amount</td>
<td>moderately, &lt;#&gt; [glasses/drinks/bottles/times]</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>occasionally, daily, rarely, on a weekly basis</td>
</tr>
</tbody>
</table>

Table 2. An example of representation of nicotine use statement.

“The patient has a history of heavy tobacco abuse for many years.”

| Status   | = history |
| Temporal | = many years |
| Amount   | = heavy   |
| Abuse    | = evidence of dependence |

Extracting substance use information from clinical texts with NLP

Natural language processing (NLP) technologies have been used for extracting a wide range of information (e.g., drugs, diseases, and findings) from clinical notes (4, 18-20). Currently, most established clinical natural language processing systems (e.g., MedLEE(19) and cTAKES(21)) as well as clinical annotated corpus (e.g., MiPACQ(22)) are primarily focused on extracting named entities such as Unified Medical Language System (UMLS) concepts (e.g., diseases, medications, or procedures). A number of efforts have focused on using automated NLP techniques to extract smoking status (e.g., “Past smoker,” “Current smoker,” and “Smoker”) and assess clinician adherence of tobacco treatment guideline from clinical notes such as discharge summaries (23-29). In Uzuner’s work, the authors described several systems for classifying the smoking status of the patients by using machine learning and rule based algorithms. These systems presented in the paper reported F-scores from 84 – 90. In this study, our preliminary classification results showed that a rule-based classifier could achieve F-scores from 85.9 – 95.9. For extracting of deeper information related to substance usage other than status such as amount, type, there has been limited work on collecting these detailed information on substance use more comprehensively.

We previously incorporated a family history model into our open-source clinical NLP system, BioMedICUS(30), for automated extraction of family history information that identifies observations (e.g., disease or procedure), relative or side of family with the attributes (e.g., vital status, age of diagnosis, certainty, and negation) and predication. In this study, we sought to extend BioMedICUS for automated detection of substance use statements and extraction of detailed information within each statement. We aimed to utilize the patterns and lexicon collected from the multi-site sources used in previous work (14) along with the deep dependency relationships between tokens within statements provided by the Stanford Dependency parser (31) to extract elements from social history statements. To augment our system performance, we also leveraged semantic labels in the Propbank corpus and the MiPACQ corpus.

Methods

Corpora and annotation

The corpora used in this study included 491 “Consult - History and Physical” notes from MTSamples.com (MTS), a public web repository with about 5,000 sample clinical notes (“Development Corpus”), and 200 “HP” notes from the University of Pittsburgh Medical Center (UPMC) de-identified clinical notes repository (32) (“Evaluation Hold-out Corpus”). The final corpora included clinical notes in a range of different specialties from both acute and ambulatory settings. The corpora annotation consisted of three rounds of annotation. First, the collected clinical notes were annotated for the social history contents as well as social history section headers with the General Architecture for Text Engineering (GATE)(33), a Java framework for developing NLP pipelines. To establish
agreement regarding annotations, a subset of 100 notes were studied collectively by TJW and EWC prior to annotating the remaining set of annotations. In the next step, annotated social history text of each note was extracted from the GATE XML output and categorized into different sets of statement types (e.g., “ALCOHOL_USE,” “OCCUPATION,” and “DRUG_USE”) with same annotation tool. Finally, sentences of three social history statement types, “ALCOHOL_USE,” “DRUG_USE,” and “TOBACCO_USE” were further annotated for elements (e.g., amount, frequency, and temporal) as well as the relationships between those elements with the brat rapid annotation tool (BRAT)(34), a web-based tool for structured annotation. Table 3 shows a brief description of the six elements that were extracted from our annotation guidelines for substance use.

**Table 3. Description and examples of elements for alcohol, drug and nicotine use.**

<table>
<thead>
<tr>
<th>Element</th>
<th>Brief Description</th>
<th>Alcohol Use</th>
<th>Drug Use</th>
<th>Nicotine Use and Exposure</th>
</tr>
</thead>
</table>
| Amount        | How much a substance used | • Few glasses wine/day  
• beer or wine 1-2 per month | • <1 joint QOD for pain  
• Daily 1 pill 4X/day | • 1-2 cigarettes/day  
• heavy tobacco use |
| Status        | Current or past substance use. | • quit many years ago  
• current drinker | • clean since 2004  
• History of drug abuse | • Minimal smoker  
• Current tobacco |
| Type          | Type of substance use or exposure | • Few glasses wine/day  
• beer or wine 1-2 per month | • Used OxyContin prescribed to mother  
• Overdose of Molly | • Occasional Cigar  
• Smoked pipes and cigars |
| Frequency     | How often substance used | • one every 2 months  
• Occasional beer | • very rare  
• Marijuana 2x/week | • 1-2 cigarettes/day  
• a pack per week |
| Method        | How substance used or how exposed | • drinks wine | • Injected heroin for years  
• Occasional IVDU | • smokes a couple  
• secondhand smoke exposure |
| Temporal      | Temporal information e.g., date started, age started, date quit and duration of use | • started drinking in 1985  
• drink since age 20  
• quit recently  
• drink for 20 years | • First joint in 1987  
• methadone 2007 stopped  
• used 10 years | • started smoking in 1985  
• quit years ago  
• Smoked for 4-5 yrs |

Figure 1 shows the annotation of a set of alcohol use sentences using BRAT. For each social history statement type, 10% of the sentences of each statement type were annotated by two annotators for inter-rater agreement on the annotation of elements. Four informatics experts provided manual annotations, including a physician, one biomedical informatics PhD, and two biomedical informatics graduate students.

![Figure 1. Annotation of “ALCOHOL_USE” sentences in clinical notes. Green shaded box represent elements and arrows represent relationships between elements.](image-url)
**Substance use statement detection**

Figure 2 shows the overview of the substance use statement detection module development. The substance use statement detection rules were built with half of the 246 MTS “Consult - History and Physical” notes. Another two sets of MTS notes were used as a development set (N=123) and a holdout test set (N=122).

All UPMC “H&P” notes were used as a separate evaluation set. Text was extracted and split into sentences by a sentence splitter. All tokens within were normalized based on the SPECIALIST Lexicon (35) and substance use sentences were extracted.

A substance use detection lexicon was collected from all our previous work on social history and substance use as key features for substance use statement detection (13, 14, 36). Overall, three lexicons for alcohol, drug, and nicotine were created from MTSamples/FHS/UVMMC notes, UVMCC comments, and HL7 Social History Observation and openEHR alcohol related archetypes. Section headers and sub-headers of each statement were also used as a feature for substance use statement detection. Section header identification was aided by a limited set of header expressions, capitalization, spacing, and punctuation header patterns. A set of regular expression based rules for substance use statement classification was manually developed using the lexicons, section headers, and iterative error analysis on the training dataset to improve the system performance. In each iteration, rules were modified and new rules were included to capture cues from section headers or sentence text. For example, if a sentence contains any family member, then the sentence was not classified as a substance usage statement. In total 10 rules were created for substance classification. The developed substance use statement detection system was evaluated on two evaluation sets: MTS test set notes (N=122) and UPMC evaluation notes (N=200).

In this study we also implemented a support vector machine (SVM) model for alcohol usage statement detection. Latent dirichlet allocation (LDA) and information gain (IG) were used for feature selection. Selected features are then used by supervised SVM machine learning to classify each statement in the training dataset. LDA with Gibbs Sampling (iteration = 1500) was implemented in Stanford Topic Model Toolkit (TMT-0.4.0) (37). Topic numbers range from 50–800 were chosen. Keywords for each topic were obtained from the output files. We then used IG to rank those keywords for each topic and chose filtered top numbers of keywords as features to implement classification.

**Elements extraction**

As shown in Table 4, the number of occurrences for some elements (e.g., amount and frequency) was limited in the small MTS development corpus. Thus, the attribute extraction module was developed on all MTSamples annotated
statements for alcohol (N=243), drug (N=130), and nicotine (N=311) use for the six elements (e.g., amount, frequency, and temporal). The UPMC annotated statements were used as the single holdout test set for evaluation of element extraction. Annotated sentences were first parsed by Stanford Parser(38) for constituent and dependency parses. The constituent parse of a sentence provides syntactic cues for elements extraction. For example, a phrasal phrase with a time period token (e.g., year and month) could be a temporal element (e.g., “since this year”). Based on this fact, a set of rules was built to detect for temporal elements by using constituent parses. For instance, the constituent parses provided prepositional phrases which were helpful for temporal elements such that the pattern “(PP (IN (for|until|since|in|on|at))” and with a time marker could be used for temporal element detection. The dependency parse of a sentence can provide the dependency structure of the sentence. A dependency structure represents a directed graph between the tokens of a sentence, where edges denote pairwise grammatical relationships (e.g., determiner [the – patient] or adjectival modifier [significant - modified]).

Figure 3 shows an example dependency structure for “She denies any significant tobacco or alcohol...” The dependency structure of a sentence captures relationships between elements and the substance (e.g., alcohol or drug). These relationships can help to extract only related elements for a complicated sentence or sentences that with multiple substances. For instance, the dependency structure for “She is a smoker about a pack and a half for 38 years and notes rare alcohol use.” can help to detect the right substance, smoking presumably tobacco or alcohol in this case, and the amount phrase “about a pack and a half for 38 years” is related to smoking. In our training dataset, we observed that for sentences like “No history of tobacco, alcohol, or illicit drug use.,” which occurred frequently in clinical notes, the distance based algorithms is obviously will not be able to detect the correct relationship between entities. On the other hand, a dependency parse can easily detect such a relationship. In this study, we used Stanford Parser v3.5.1 to generate dependency structures for social history statements.

For temporal extraction, we observed that temporal elements varied largely in terms of syntactic patterns and lexicon (e.g., “most of her adult life,” “up until about ten years ago,” “for a while”). The syntactic patterns and lexicons collected from our previous studies were not adequate to extract temporal elements. In addition to the lexicon collected from previous study, we used the temporal semantic role (ARG-TMP) annotations from Propbank(39) and MiPACQ(40) – the latter of which is a clinical corpus annotated with PropBank-style predicate argument structures. The Propbank and MiPACQ are two large corpora annotated with semantic roles (e.g., goal, temporal, manner, purpose) of all the phrases in a sentence. This semantically annotated sentence shows an example of how the temporal role annotation on constituent parse which includes temporal expressions “TMP”: “(S (NP-SBJ (DT The) (NN patient) ) (VP (MD will) (VP (VB be) (ADJP-PRD (JJ ready) (NP-TMP (NNP Feb.) (CD 15) )))) (.) )”. Temporal lexicons and patterns collected from the temporal semantic role annotations in these two resources were included into the rules for temporal elements detection. For example, temporal phrases like “at the same time” and “yesterday” were included as extra lexicon. The resulting new syntactic patterns captured additional temporal patterns including many that are not in the training dataset or the original lexicon and had improved performance. For each element of the three types of substance use statements, a set of regular expression heuristic rules were developed based on the lexicon, syntactic structure, and dependency structure. The created module was tested on substance use statements annotated from 200 UPMC notes.

In order to extract all 6 elements, each alcohol, drug and nicotine statement was first searched for patterns (e.g., “up to 6 drinks”, “3-packs”) and lexical items (e.g. “significant”, “dips”), represented by regular expressions. Then, the constituent parse of the statement was searched for syntactic patterns of each element (e.g., “(PP IN until)”. The searched phrases were validated by dependency parse of each statement. The patterns, lexicon and syntactic patterns used for elements extraction were collected from multi-site sources in our previous work14, MiPACQ and Propbank, with a number range from 6 to more than 3,000. A set of rules was manually created to determine if a searched phrase is related to the particular substance of interest by using iterative error analysis on the training dataset. In each iteration, rules were modified and new rules were included to accurately detect the correct relationships. Below are two examples of such rules. Some element types often needed more rules (e.g., 15 rules for temporal) to validate
which substance was associated with a searched phrase. On average, 5 rules (median 3, range 1-15) were created for each element extraction task.

Example 1). IF all parents of tokens in the searched phrase consist of the drug token AND the searched phrase is not negated THEN the searched phrase is associated with the drug token.

Example 2). IF the statement contains only the drug token without alcohol or nicotine keywords THEN the searched phrase is associated with the drug token.

Results

Table 4 summarizes the statistics of overall annotations of the corpora. From 491 MTS notes, 234 sentences were marked as alcohol use related, 124 were drug use related, and 260 were nicotine use related. From UPMC notes, 138, 72 and 148 sentences were marked as related to alcohol, drug and nicotine use. The number of occurrence of each element with these sentences is also listed in Table 4.

Table 4. Summary of corpora annotations.

<table>
<thead>
<tr>
<th>(A) Substance use (SU) text annotation</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Note type</td>
<td>No. of notes</td>
<td>No. of notes with SU</td>
</tr>
<tr>
<td>MTS</td>
<td>Consult – H&amp;P</td>
<td>491</td>
<td>378</td>
</tr>
<tr>
<td>UPMC</td>
<td>H&amp;P</td>
<td>200</td>
<td>179</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(B) Substance use statement type annotation</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Statement type</td>
<td>No. of notes with statement type</td>
<td>No. of sentences</td>
</tr>
<tr>
<td>MTS</td>
<td>Alcohol use</td>
<td>234</td>
<td>243</td>
</tr>
<tr>
<td></td>
<td>Drug use</td>
<td>124</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>Nicotine use</td>
<td>260</td>
<td>311</td>
</tr>
<tr>
<td>UPMC</td>
<td>Alcohol use</td>
<td>138</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Drug use</td>
<td>72</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Nicotine use</td>
<td>148</td>
<td>172</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(C) Substance use elements annotation</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Statement type</td>
<td>No. of Amount occurrence</td>
<td>No. of Status occurrence</td>
<td>No. of Type occurrence</td>
<td>No. of Frequency occurrence</td>
</tr>
<tr>
<td>MTS</td>
<td>Alcohol use</td>
<td>31</td>
<td>44</td>
<td>193</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Drug use</td>
<td>13</td>
<td>34</td>
<td>148</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Nicotine use</td>
<td>62</td>
<td>144</td>
<td>124</td>
<td>53</td>
</tr>
<tr>
<td>UPMC</td>
<td>Alcohol use</td>
<td>19</td>
<td>36</td>
<td>125</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Drug use</td>
<td>20</td>
<td>17</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Nicotine use</td>
<td>25</td>
<td>90</td>
<td>48</td>
<td>21</td>
</tr>
</tbody>
</table>

The Cohen’s kappa inter-rater agreement between two annotators for elements of alcohol, drug, and nicotine use statements was 83.4, 80.9, and 90.6. Table 5 shows substance use statement detection performance on the MTS test set (N=122) and UPMC test set (N=200). As summarized, detection of alcohol, drug, and nicotine use statements was good with high F-scores.

Table 6 summarizes the performance of the system for extraction of substance use elements on UPMC substance use statements. The type, method and amount detection showed good performance based on F-scores for alcohol, drug and nicotine use sentences. Frequency and status detection for drug and nicotine use showed better performance for drug use statements and for nicotine use statements compared to alcohol use statements.

After inclusion of extra patterns and lexicon from general English semantic resources and clinical annotated semantic resources, temporal detection improved greatly, though this was still not as good as the performance of other elements, primarily due to the large variety of potential temporal expressions. Figure 4 shows the improvement of temporal element extraction with alcohol use and drug use statements achieved by adding in outside lexicons and patterns from MiPACQ and Propbank.
Table 5. Substance use statement detection performance.

<table>
<thead>
<tr>
<th>Substance use</th>
<th>Sensitivity</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTS test set</td>
<td>91.8</td>
<td>97.8</td>
<td>94.7</td>
</tr>
<tr>
<td>UPMC test set</td>
<td>83.1</td>
<td>97.6</td>
<td>89.8</td>
</tr>
<tr>
<td>Drug use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTS test set</td>
<td>84.0</td>
<td>87.5</td>
<td>85.7</td>
</tr>
<tr>
<td>UPMC test set</td>
<td>76.0</td>
<td>98.3</td>
<td>85.7</td>
</tr>
<tr>
<td>Nicotine use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTS test set</td>
<td>96.6</td>
<td>95.0</td>
<td>95.8</td>
</tr>
<tr>
<td>UPMC test set</td>
<td>89.1</td>
<td>89.7</td>
<td>89.4</td>
</tr>
</tbody>
</table>

Table 6. NLP system element extraction performance for substance use statements.

<table>
<thead>
<tr>
<th>Element</th>
<th>Sensitivity</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount</td>
<td>68.0</td>
<td>94.4</td>
<td>79.1</td>
</tr>
<tr>
<td>Frequency</td>
<td>73.3</td>
<td>73.3</td>
<td>73.3</td>
</tr>
<tr>
<td>Status</td>
<td>75.0</td>
<td>61.4</td>
<td>67.5</td>
</tr>
<tr>
<td>Method</td>
<td>84.6</td>
<td>80.5</td>
<td>82.5</td>
</tr>
<tr>
<td>Type</td>
<td>98.4</td>
<td>87.2</td>
<td>92.5</td>
</tr>
<tr>
<td>Temporal</td>
<td>61.9</td>
<td>68.4</td>
<td>65.0</td>
</tr>
<tr>
<td>Drug use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount</td>
<td>77.3</td>
<td>94.4</td>
<td>85.0</td>
</tr>
<tr>
<td>Frequency</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Status</td>
<td>83.3</td>
<td>93.8</td>
<td>88.2</td>
</tr>
<tr>
<td>Method</td>
<td>76.9</td>
<td>100.0</td>
<td>87.0</td>
</tr>
<tr>
<td>Type</td>
<td>98.9</td>
<td>100.0</td>
<td>99.4</td>
</tr>
<tr>
<td>Temporal</td>
<td>66.7</td>
<td>100.0</td>
<td>80.0</td>
</tr>
<tr>
<td>Nicotine use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount</td>
<td>84.0</td>
<td>80.8</td>
<td>82.4</td>
</tr>
<tr>
<td>Frequency</td>
<td>95.2</td>
<td>100.0</td>
<td>97.6</td>
</tr>
<tr>
<td>Status</td>
<td>76.7</td>
<td>100.0</td>
<td>86.8</td>
</tr>
<tr>
<td>Method</td>
<td>93.5</td>
<td>100.0</td>
<td>96.7</td>
</tr>
<tr>
<td>Type</td>
<td>95.8</td>
<td>100.0</td>
<td>97.9</td>
</tr>
<tr>
<td>Temporal</td>
<td>74.4</td>
<td>80.0</td>
<td>78.4</td>
</tr>
</tbody>
</table>

Figure 4. Temporal element detection performance improvement.
Discussion

In this study, we developed an NLP system for substance use statement detection and element extraction from clinical notes based on previous substance use models and the addition of developed lexicons(14). The system provides functionalities to detect three types of substance use statements: alcohol, drug, and nicotine use from free-text clinical notes, as well as extraction of important semantic elements as defined in substance use models previously developed(14) including amount, frequency, type, status, method and temporal. Overall, the developed system performed well, although certain elements like temporal expressions had challenges due to the variability in expression with these items.

Similar to previous work on family history extraction from clinical texts(41), we used the dependency structure for capturing relationships between phrases and tokens in the substance use statement. We found that the dependency relationship could help attaching possible element tokens or phrases to the right substance (e.g., alcohol or drug). However, due to dependency parsing errors on long, complicated sentences, relationships were sometimes wrongly collected from the dependency structure sometimes leading to errors on attaching tokens to correct types or methods. Further study is needed for reducing this type of errors. In our experiments, we did not observe sentences that caused the parser to break. Constituent parsing errors also caused sentence-segmenting errors, which in turn affected the element detecting process.

As shown in previous work(41), rule-based approaches can achieve comparable performance as machine learning approaches. In this study we choose to implement a rule-based NLP system by leveraging features from a number of sources and previous work, particularly to improve substance use lexicons. Substance use statement detection results of the rule-based systems showed good performance on both the MTS and UPMC test sets for all three types of substance use statements. Further, our pilot experiments on alcohol statement detection using models based on topic analysis and machine learning showed no performance improvement compared with the rule-based statement detection.

We recognize as a limitation that our system would ideally have a baseline gold standard with which to evaluate itself against. Currently, gold standard corpora and comparable systems do not exist. In the future, we plan to release publically available corpora with annotations so that future systems will have a baseline corpora with which to evaluate itself.

This study was also limited in that development of statement detection and elements extraction portions of the NLP system were built on one corpus (MTS notes) and tested on a relatively small hold-out corpus (UPMC notes). While this approach suggested that our findings could be generalized to clinical notes from other sites, validation with data from another institution would be helpful in confirming these results. A next step will therefore include setting further assessments of the system on institutional clinical notes and expansion of the system for substance use comments from the structured EHR social history module. Annotations used for extracting frequency, method and temporal for drug use statements were limited in this study. Although the same approach used for extracting these same elements for alcohol use and nicotine use had good results, validation with additional drug use statements would help with confirming our findings.

Conclusion

In this study, we achieved reasonable performance for both substance use statement detection and element extraction from two clinical notes corpora. The results of this study are promising for automated detection of free-text substance use statements and extraction of detailed substance use information from clinical notes. Next steps include further refinement of element extraction for better performance, additional evaluations to demonstrate generalizability of the system, and development of modules for extraction of additional social history statements.

Acknowledgements

The National Institutes of Health through the National Library of Medicine (R01LM011364 and R01GM102282), Clinical and Translational Science Award (8UL1TR000114-02) supported this work. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.
References


33. GATE. [March 1, 2015]; Available from: https://gate.ac.uk.


Conversational Agents for Automated Inpatient and Outpatient Health Counseling

Timothy Bickmore, Ph.D.¹
¹College of Computer and Information Science, Northeastern University, Boston, MA

Abstract
Conversational agents are computer-animated characters that can simulate one-on-one, face-to-face counseling with a health provider. These agents have now been used to provide automated health education and health behavior change counseling in a wide range of medical domains. In the inpatient setting they have been used to counsel on options for surgical anesthesia, post hospital-discharge self-care, and to provide continual hospital bedside education, sleep promotion, and patient activation. Conversational agents have also been integrated with sensor systems so that they can provide context-sensitive counseling and explain events occurring in the hospital room to patients, as well as provide a mediated communication channel between patients and their providers. In the outpatient setting they have been used to collect family health histories, screen for and counsel on substance abuse, and to provide medication, exercise, diet, breastfeeding, and preconception care promotion. This medium has been shown to be particularly effective for patients with low health literacy, and has been linguistically and culturally tailored for several populations, including Latino and Chinese. Several of these systems will be demonstrated, and the development methodology and underlying technology described, including 3D animation systems, dialogue engines, and behavioral ontologies for driving automated health behavior change counseling. Evidence from trials comparing conversational agents to more conventional methods and technologies will be presented.

Specific purpose/problems
Patient behavior, such as non-adherence and poor lifestyle health behavior, is a leading cause of morbidity and mortality. These problems are especially severe in disadvantaged populations with high prevalence of inadequate health literacy. Providers do not have the time these patients require to ensure they have the understanding, confidence, and motivation to take care of themselves.

Innovation
Conversational agents combine procedural animation with dialogue systems and sensors to provide a simulated face-to-face conversation with an empathic, expert health provider.

Degree of deployment
Over 25 randomized clinical trials have been completed or are underway to evaluate virtual agents in the role of health counselors.

References
Navigating between Drug Classes and RxNorm Drugs with RxClass

Olivier Bodenreider, M.D., PhD, M.S., Lee Peters, M.S., Thang Nguyen, M.S.
National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Contact information: RXNAVINFO@LIST.NIH.GOV

Abstract: RxClass is a web-based, interactive browser and companion application programming interface (API) to explore the relationships between RxNorm drugs and drug classes from several sources including ATC, MeSH and NDF-RT. Like RxNav, RxClass is publicly available at: http://rxnav.nlm.nih.gov

System Description:

RxClass supports the following drug classification systems:

- The Anatomical Therapeutic Chemical drug classification (ATC) is a resource developed for pharmacoepidemiology purposes by the World Health Organization Collaborating Centre for Drug Statistics Methodology.
- The Medical Subject Headings (MeSH), developed by the National Library of Medicine (NLM), provides a rich description of pharmacological actions for the purpose of indexing and retrieval of biomedical articles.
- The National Drug File-Reference Terminology (NDF-RT), developed by the Department of Veterans Affairs, provides clinical information about drugs and contains the following class types: FDA Established Pharmacologic Classification (EPC), Disease classification, Chemical Structure and Classification (Chem), Mechanism of Action (MoA), Physiologic Effects (PE) and Pharmacokinetics (PK) class types.

ATC and MeSH provide both the vocabulary for drug classes and the drug-class membership relations. In contrast, several sources - DailyMed, FDASPL and NDF-RT - provide drug-class membership relations in reference to the NDF-RT vocabulary for classes. All drugs are normalized to RxNorm.

RxClass provides a graphical interface to explore the hierarchical class structures of each source and examine the corresponding RxNorm drug members for each class. Some features of RxClass:

- The user can navigate through the drug classes via the hierarchical menu (left pane), or use the search feature to identify a drug class or RxNorm drug (top search bar). RxClass contains an autocomplete function which will help identify class or drug names in search mode, as well as spelling suggestions for misspelled drug and class names during search.
- Selecting a class displays the list of its drug members in a table (main area), showing their identifiers and names in both RxNorm and the original classification system.
- RxClass supports the exploration of all classes for a given drug across multiple classifications (All/classes/Show button).
- RxClass provides a list of similar classes based on drug membership to the selected class. Users can select a class from the list and view a detailed table listing the drug members of each class with a Venn diagram highlighting the overlapping members.

Companion API. RxClass is supported by functions from an application programming interface (API), which can be used independently for integrating drug class information in programs. The API serves the latest information available from the drug information sources. The RxClass API can interoperate with the RxNorm API. The demonstration will include examples of using the API to mimic the functionality displayed in RxClass.

RxClass has been operational since July 2014. From October 2014 through June 2015 RxClass has received an average of 1254 queries per month. The drug information sources are updated monthly. The RxClass API has been operational since September 2014. RxClass supports the analysis of medication datasets with different drug classification systems, as well as the analysis of medication value sets in reference to drug classes.

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.
SEMCARE - Semantic Data Platform for Healthcare

Philipp Daumke¹, Claudia Riede¹, Thomas Fassbender¹, Ángel Honrado², Markus Kreuzthaler³, Pablo López-García³, Stefan Schulz³, Erik van Mulligen⁴, Herman van Haagen⁴, Jan Kors⁴, Hanney Gonna⁵, Xinkai Wang⁵, Elijah R Behr⁵

¹Averbis GmbH, Freiburg, Germany; ²SYNAPSE Research Management Partners S.L., Barcelona, Spain; ³Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria; ⁴Erasmus University Medical Center, Rotterdam, The Netherlands; ⁵St. George’s University of London, United Kingdom

Description

The need for exploiting medical data for secondary use has grown tremendously over the last years and covered in various research initiatives such as EHR4CR¹, i2b² and eMERGE³. Aggregated patient-level data can support the identification of disease mechanisms and new discovery areas, improve drug safety surveillance and decrease patient recruitment cycle times for clinical trials. Exploiting patient-level data can optimize clinical studies in several ways, e.g., by enabling the definition of appropriate study design or ensuring that inclusion/exclusion criteria map to an existing patient population. As large parts of patient-level data in electronic health records (EHRs) are only available as free text, language technologies are an indispensable prerequisite for this process. SEMCARE is an EU-funded project that is building a semantic data platform able to define patient cohorts based on clinical information scattered in heterogeneous resources. Three hospitals from the Netherlands, UK and Austria serve as pilot sites, focusing on cardiology use cases. However, SEMCARE’s long-term objective is to build a flexible information extraction and semantic indexing platform that can be adapted to a broad range of languages and clinical contexts.

Figure 1 shows the different modules of the SEMCARE platform. Data integration deals with the management and transformation of heterogeneous data (e.g., ASCII, Word, PDF, XML or HL7) into homogeneous formats. State-of-the-art text mining technologies⁴ based on the framework Apache UIMA⁵ and multilingual semantic resources (e.g. UMLS subsets, ICD10, LOINC, ATC) are used to extract information from EHRs. The results are fed into a hybrid free-text and concept-based search engine. The anonymization service⁶ allows text de-identification to ensure patient privacy and to make clinical data accessible to a broader research audience.

In the demo session, we will present the semantic data platform developed in SEMCARE based on anonymized clinical data. We will show how this platform is able to identify patient cohorts based on patient-level criteria, much of which are found in weakly structured text. In addition, we will present the clinical evaluation of the platform that is currently tested by three major European hospitals in a cardiology use case.

Figure 1. SEMCARE Architecture

5 https://uima.apache.org/
Leveraging Health Information Exchange to Create Neighborhood Health Records for Public Health Agencies

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

Abstract
Assessment is a core function of public health. Traditional methods leverage surveys, vital records, and paper-based disease reporting. Increased adoption of electronic health records and health information exchange present an opportunity to enhance health assessment activities by enabling access to up-to-date data on disease burden and outcomes, with sufficient geographic density to allow small area analyses. Furthermore, when combined with community-level data relevant to social determinants, HIE has the capacity to create a neighborhood health profile that measures and tracks health risks and disparities at a community level. In this demonstration, we will present how we’ve leveraged routinely collected and geospatially-enhanced EHR data available from a regional health information exchange to create neighborhood health profiles for a county health department. The discussion will highlight neighborhood level indicators developed for the project, data visualizations, and issues of representativeness or bias in the data.

Introduction
Public health authorities monitor population health to identify burden of disease, manage health assets, establish policy, and evaluate interventions. This assessment usually relies on information available through surveys, vital records, and paper-based disease reporting. These sources capture sparse data at the community level. Electronic health record (EHR) systems may provide more timely and dense data for geographic areas of interest. Yet there exist a number of challenges to routine use of EHR data, including linking them to social determinant data, and adjusting them so they reflect the general population rather than those most likely to use health care services.

Description of the System Demonstration
In this demonstration, we will describe how we developed and implemented neighborhood health profiles for use by a local health department. We will show the profiles created and discuss how we calculated health indicators using EHR data. Furthermore, we will describe the tools (e.g., Web-based Analysis and Visualization Environment: Weave) used to enable custom visualizations of the health indicator data joined with population level data.

Neighborhood health profiles consist of community-level indicators using EHR data integrated with a community information system (CIS). These indicators represent geographic units smaller than a county, the traditional unit extrapolated from large population-based surveys and datasets. The neighborhood health profiles were created using a large health information exchange with over 5 billion clinical observations from heterogeneous EHR systems. Patient level data were geocoded and aggregated then integrated with data from the CIS, which possesses data on transportation, crime, education, poverty, and other social determinants of health. Neighborhood level indicators include the prevalence of diseases of public health interest (e.g., diabetes mellitus type 2, depression, chlamydia infection) and several HEDIS-like clinical quality indicators (e.g., number of eligible patients screened for chlamydia, number of diabetics receiving annual HbA1c testing). Indicators were calculated at various levels of geographic granularity (e.g., ZIP code, census tract, census block group, neighborhood, city-county council district).

Once generated, neighborhood health profiles were shared with the local health department for review and feedback, and were compared to results from a random population sample survey. Geospatial visualization as well as statistical models were developed in partnership with the health department to aid interpretation and application to public health practice as well as policy. Such uses for EHR and HIE infrastructure are novel and address gaps identified by recent Institute of Medicine as well as Robert Wood Johnson Foundation expert panels which published recommendations for strengthening public health surveillance and research.
System Description and Areas of Innovation

The Observational Health Data Sciences and Informatics collaborative (OHDSI, pronounced ‘Odyssey’) was formed in 2013 with the goal of creating reliable scientific evidence through large-scale analytics of observational health data. A cornerstone of OHDSI is the use of a common data model (CDM) by all participants. Specifically, OHDSI leverages a CDM that was first developed in 2008 as part of the Observational Medical Outcomes Partnership and is now in its 5th iteration.

In this demonstration, we will be presenting open-source tools from the OHDSI software ecosystem. All source code for these tools is available on Github (www.github.com/ohdsi) and live demos can be found on the OHDSI website (www.ohdsi.org/demos). We will focus on several key innovations essential for collaborative data analytics. First, we will introduce software resources for transforming diverse datasets into the CDM, including an automated dataset mapper and native support for multiple database dialects (e.g., SQL Server, Oracle, Postgres, Netezza).

We will then take the audience through a clinical use-case, introducing a series of web-based tools that leverage the CDM for dataset characterization and quality assessment, patient cohort development, and data analytics and visualization (Figure 1). We will also present a novel tool for exploring multiple standardized vocabularies, a cornerstone for creating cohorts and performing large-scale observational studies. The OHDSI visualization library, based on D3 (www.d3.js), will be described in conjunction with these tools.

For the final section of the demonstration, we will return to OHDSI’s core objectives of evidence generation and dissemination. We will demonstrate the use of the OHDSI research platform, which leverages the R statistical package to disseminate and customize analyses for individual study sites. We will also highlight the OHDSI pharmacovigilance API, a REST web service for retrieving evidence on adverse drug events from a wide variety of sources, including product labeling, adverse event reporting, the scientific literature, clinical registries, and other sources.

Current Deployment: As of October 2014, 58 distinct databases had been converted or were in the process of being transformed into the OMOP CDM. As of March 2015, OHDSI tools are in use at more than 10 sites, including Columbia, Stanford, Regenstrief Institute, University of Pittsburgh, National Institutes of Health, Ajou University, Erasmus MC, Janssen Research and Development, and AstraZeneca.
Towards an Open EHR Platform: Porting a Complex Application using SMART on FHIR
Alistair Erskine, MD1; David P McCallie, MD2; Marcy Stoots DNP, RN-BC3
1Geisinger Health System, Danville PA; 2Cerner Corporate, Kansas City MO; 3xG Health Solutions, Columbia MD

Abstract
The Rheumatology App is a sophisticated, web-based application for managing a rheumatology ambulatory encounter, developed by Geisinger Health System. The Rheumatology App synthesizes information from several sources, including the EHR and presenting disease-specific data in a graphical summary view. It assists in the formatting of a progress note for the visit, incorporating relevant data from all sources. The Rheumatology App was initially developed to run on a major commercial EHR (Epic). Subsequently, it was adapted to take advantage of the SHARP grant funded SMART platform and emerging HL-7 FHIR resources, given that these draft standards are publicly available, non-proprietary and facilitate the exchange of clinical data in an EHR-agnostic manner. The data set used for these RESTful APIs were derived from the Meaningful Use Common Dataset, as promoted by the Argonaut Project and the Office of the National Coordinator (e.g. LOINC, RxNORM, SNOMED-CT). The process of converting the Rheumatology App from an embedded state within one commercially available EHR into a “SMART on FHIR” state into another EHR was shorter than expected, measured in weeks, not months or years. Converting the full functionality required at least eleven (11) FHIR resource types, some of which were both read and write access to the EHR. This is one of the more complex implementations using SMART on FHIR to appear in the literature to date. We will demonstrate the Rheumatology App executing as a “SMART app” in the context a commercial EHR. We will discuss the process, the learning from the conversion exercise and deploying the Rheumatology App as SMART on FHIR functionality embedded directly and unobtrusively within the framework and workflow of the various EHRs, from the perspective of both the health system application developers, and the EHR vendor.

Purpose of the System
The Rheumatology App presents clinical information in a format that supports the cognitive process of a clinician treating rheumatic disease. All Geisinger Rheumatologist use this Rheumatology App to manage their patients. The workflow includes touchscreen questionnaires for the patient, designed to collect information about symptoms of their new or chronic rheumatologic condition. The patient generated data is rendered into views that, when assembled with data imported from the EHR and data from clinician input, facilitate shared decision making and visualization of the next best actions to perform for the patient. Following the Learning Health System principles, as a by-product care administered, this data entry and integration supports the development of a rheumatic database for research and decision support. In addition, increases in clinician efficiency and productivity were measured using the Rheumatology App. The SMART on FHIR implementation of the Rheumatology App demonstrates the feasibility of delivering complex and specialized clinical functionality using a standards-based open platform approach.

Deployment Status
In its native implementation, the Rheumatology App has been used in production for several years, and seen widespread adoption among eligible clinicians. The SMART-on-FHIR version of the APP is currently running as a proof-of-concept. Below is a screenshot of the Rheumatology App executing in a commercial EHR.
RapTAT: A Tool for Assisted Annotation and Reviewer Training via Online Machine Learning

Glenn T. Gobbel, DVM, PhD, MS\textsuperscript{1,2}, Ruth M. Reeves, PhD\textsuperscript{1}, Brett R. South, MS, PhD\textsuperscript{3}, Wendy W. Chapman, PhD\textsuperscript{3}, Jay Jarman, PhD\textsuperscript{4}, Steven Lay, BS\textsuperscript{4}, Michael E. Matheny, MD, MS, MPH\textsuperscript{1,2}

\textsuperscript{1}Tennessee Valley Health System VA, Nashville, TN; \textsuperscript{2}Vanderbilt University, Nashville, TN; \textsuperscript{3}University of Utah, SLC, UT; \textsuperscript{4}East Tennessee State University, Johnson City, TN

Abstract
Manually annotating clinical free text to extract the valuable information it contains is time-consuming and expensive. The task commonly relies on medically-trained human reviewers to identify concepts related to treatment and care. Although natural language processing (NLP) systems can partially automate the data extraction process and thus reduce the cost, such systems still depend on manual annotations for system training and testing. We have developed the Rapid Text Annotation Tool (RapTAT), the goal of which is to reduce the burden of free text review by assisting with both annotator training and the annotation process itself. Training of the tool relies on real-time online learning to identify the phrases and concepts of interest to the annotator, and it employs this training to pre-annotate subsequent documents, thus reducing the time required to find and annotate phrases of interest. The tool also assists with annotator training by providing visual feedback that compares and contrasts annotations of inexperienced or non-expert reviewers to those provided by a reference standard.

Description
The Rapid Text Annotation tool (RapTAT) currently consists of two modules: 1) a reviewer versus reference (RVR) module and 2) a pre-annotation (PA) module, both of which are incorporated into an existing annotation tool, eHOST.\textsuperscript{1} During annotator training, the RVR module provides feedback to inexperienced reviewers by calculating performance scores relative to the reference standard (F-measures) for each concept of interest; it also provides visual cues highlighting differences between the annotations of the trainee and an expert-generated reference standard. The tool tracks changes in performance score for each concept of interest versus the number of documents annotated to help users assess the benefits of further reviewer training. Once an annotator is trained, the PA module uses real-time online machine learning based on both the reference standard and information provided by ongoing reviewer annotations to pre-annotate concepts of interest before document review. The tool does not rely on a manually-generated dictionary for pre-annotation; instead, it uses natural language processing (NLP) dynamically to model the annotations of the reviewer. The NLP uses a novel probabilistic model for identifying phrases for annotation and a token -order-specific naïve Bayes model for concept mapping. It continuously updates and uses this model to identify concepts for annotation within subsequent documents. As the annotation process and tool training continues, PA module pre-annotations more closely match those of the reviewer, which gradually shifts the annotation task from one of searching and highlighting phrases of interest to one focused on review and correction. This approach to pre-annotation has been used to annotate concepts predictive of acute kidney injury and pressure ulcer development and for determining document quality; it can reduce reviewer annotation time by up to 50%.\textsuperscript{2}

Deployment
The tool has been deployed to two sites in addition to the site of tool development. A study underway will assess the impact of tool use on the speed and accuracy of annotations of both medical experts and those without medical training. The results of that study should establish whether the tool can effectively reduce the time spent on annotation and reviewer training.

References
The Scalable Collaborative Infrastructure for a Learning Health System: Facilitating Agile Comparative Effectiveness Research

Jeffrey G. Klann, PhD,1,2,3, Marc Natter, MD1,4, Doug MacFadden, MS1, Sarah Weiler, PhD1, Kenneth Mandl, MD, MPH1,4, Shawn N. Murphy, MD, PhD1,2,3

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

Abstract

The Scalable Collaborative Infrastructure for a Learning Health System (SCILHS, pronounced “skills”) is a growing network of health centers across the United States. SCILHS is a Clinical Data Research Network (CDRN) in the Patient-Centered Outcomes Research Institute’s PCORnet, a national effort to instantiate a ‘network of networks’ that supports large-scale comparative effectiveness research. SCILHS uses Informatics for Integrating Biology and the Bedside (i2b2) as its technical backbone and has adopted the PCORnet Common Data Model (CDM) as its foundation for interoperable data exchange. We developed an i2b2 ontology that represents the CDM, and we developed extensive documentation and spreadsheets to assist the mapping process. We made this ontology publically available, and portions of it are used by several other CDRNs. SCILHS uses the Shared Health Research Information Network (SHRINE) platform for distributed querying. SCILHS sites have developed mutual trust relationships, which enable the SCILHS SHRINE hub to perform live queries across all sites and return aggregate counts in real-time. This prep-to-research functionality enables SCILHS to rapidly identify patient cohorts with specific conditions who meet eligibility criteria for observational research or clinical trials. SCILHS is a successful demonstration of an approach for live, distributed queries across diverse environments with disparate data.

Overview

Nationwide networks of hospital systems’ clinical data are emerging to enable large-scale comparative effectiveness research (CER). To facilitate agile CER, we find two often-overlooked elements are very important:

- **A low-friction methodology that supports participation of many sites, with data in many different formats.** Informatics expertise and technological infrastructure vary widely across hospital systems. Therefore, to attract a diverse group of sites, it is necessary to provide a well-documented methodology to support a variety of data formats.

- **Live querying on limited query types, to enable rapid identification of patient cohorts with specific conditions.** The prep-to-research activity of cohort identification is most effective if results are available without delay. As an example, consider determining if a network has enough patients to support a particular clinical trial. This query is not human subjects research and therefore with proper agreements and supportive technology, the result could be made available without personnel at each site signing off.

SCILHS includes these elements by leveraging previous successes developed in the Shared Health Research Informatics Network (SHRINE) platform. SHRINE is a distributed query protocol for i2b2 that is used in several active networks nationwide to provide live cohort-identification queries. SCILHS is unique among SHRINE networks in that:

- **SCILHS has enhanced SHRINE** to support more complex queries and to allow the patients found in these prep-to-research queries to flow into disease-specific research data marts for further analysis.

- **SCILHS has developed a new methodology for supporting semantically interoperable querying without modifications to underlying data.** This new methodology is easier for sites to implement, directly supports data transformations, and improves query speed over previous SHRINE approaches.

- **SCILHS represents the most diverse SHRINE network that is actively used for clinical research,** in terms of the governance and infrastructure of its sites, which required regulatory innovations.

Description of System

As of this writing, SCILHS is live at 7 sites in four states, covering over 8 million patients. Seven more sites are expected to be live by Fall 2015. Live cohort-finding queries can be run on all of this data, and we have begun characterizing specific diseases and generating disease-specific datamarts. Here we will demonstrate queries running at the SCILHS hub, i2b2 implementation of the PCORnet CDM, and the methodology that has allowed us to be a network that supports live prep-to-research queries. Queries are built in the SCILHS Common Data Model using the graphical SHRINE query tool. Queries are automatically sent to individual sites through the SHRINE Adapter. Each site’s i2b2 runs the query and returns aggregate counts to the SCILHS hub, which obfuscates the counts and displays the results. Attacks on the network to gain access to identifiable data through the SHRINE are impossible by design, because the queries are transmitted in an XML format that limits query complexity.
Electronic Health Management Platform (eHMP): The Next Phase of VA’s EHR

Jonathan R. Nebeker MS MD*† W. Paul Nichol MD* Shane McNamee MD* Jessica Murphy MD* James Hellewell MD* Reese Omizo MD* Kristian Johnson RN* Kensaku Kawamoto MD* Margaret McDonald MD* Theresa Cullen MD* Elaine Hunolt* Scott Woods PhD* Jennifer Herout PhD* Charlene Weir PhD*. *Dept. of Veterans Affairs; †University of Utah;

Abstract

VA is in the second year of its EHR modernization program called VistA Evolution. The Enterprise Health Management Platform (eHMP) is next generation EHR system. eHMP incrementally delivers new capabilities necessary for new models of healthcare that we desire to implement, while it preserves the business processes that we desire to keep. Technically, eHMP builds upon VistA, our current MUMPS system, with a data aggregation layer, service layer using modern open-source components, and a modular user interface that allows addition of compliant applications from outside VA. New data elements will be natively stored using national standards. Many new services use standard implementations and interfaces such as OpenCDS and Business Process Management Notation. This demonstration will highlight the ability of this system to collect and operate on data from 130 different VistA EHRs, DoD, and community partners participating in Connect or Direct data exchanges. It will also highlight the new user experience that is heavily informed by social and cognitive psychology.

Description

Enterprise Health Management Platform (eHMP) is a general-purpose EHR system. It addresses both technical and functional problems with VistA and many other EHRs. By the time of the presentation, it will be deployed throughout most of the VA system.

This demonstration will briefly highlight eHMP’s technical solution to the interoperability problem. Data collection is accomplished through remote procedure calls from 130 different instances of VistA worldwide, national health exchanges, and a special pipeline to DoD data. Incoming data is parsed and standardized, currently using maps, and stored as JSON objects. These objects are called by services and the user interface. An older version of this technology is in use at DoD. Together these technologies allow service members and Veterans to have a seamless medical record from acute treatment in Germany to ongoing treatment between Walter Reed and area VA medical centers.

The demonstration will also highlight eHMP’s innovative user experience. eHMP integrates traditionally segregated data (laboratory results, medications, problems, notes, etc) into workspaces. Data integration promotes contextual awareness and easy navigation, as users focus in and out on data during their workflows. Benefiting from standardized data, the user experience summarizes most displays by concept such as a medication ingredient or diagnosis instead of displaying individual records. Summarization at the top level of the user interface allows scanning and analysis behaviors. Consistent drill-down capabilities support quick-look, inspection, and browsing behaviors. These behaviors have analogs in mind-set theory and related theories about human understanding and action. Another innovation is incorporation of patient-tailored goals. These are editable on a patient level and are more prominently displayed instead of traditional reference ranges. Finally, eHMP also allows users to manipulate their information environments to explore a temporary question or to customize their environment to support specific workflows. Current capability focuses on diagnosis-driven workflows.
PhEMA: Phenotype Modeling, Sharing and Execution Architecture

Jyotishman Pathak, PhD\(^1\); Joshua C. Denny, MD, MS\(^2\); William Thompson, PhD\(^3\); Luke V. Rasmussen, BS\(^4\)

\(^1\)Department of Health Sciences Research, Mayo Clinic, Rochester, MN
\(^2\)Department of Biomedical Informatics, Vanderbilt University, Nashville, TN
\(^3\)Center for Biomedical Research Informatics, NorthShore University, Evanston, IL
\(^4\)Department of Preventive Medicine, Northwestern University, Chicago, IL

Abstract. The phenotype modeling, sharing and execution architecture (PhEMA) platform provides a comprehensive solution for standards-based phenotype algorithm authoring, integration with the Phenotype Knowledge Base website (PheKB, \url{http://phekb.org}) for community-wide sharing, and execution on standardized and normalized EHR data using advanced workflow management via the open-source Konstanz Information Miner (KNIME, \url{http://knime.org}). In this system demonstration, we will highlight the key features of the platform and provide hands-on experience in authoring, sharing and executing phenotype algorithms using de-identified, standardized electronic health record (EHR) data. The platform is available at: \url{http://projectphema.org}.

Background and System Description. EHRs can be used for many different types of research including disease-based, response to treatment, clinical biomarkers, redefining “normal”, and analysis of changes over time of clinical variables and parameters. Deriving such phenotypes from EHR data can be challenging, including the ability and process for authoring and executing phenotyping algorithms\(^1\). The PhEMA platform attempts to address these challenges by providing a comprehensive solution for standards-based authoring, sharing and execution of phenotyping algorithms\(^2\). More specifically, we use the National Quality Forum (NQF) Quality Data Model (QDM2) and HL7 Health Quality Measures Format (HQMF) for standardized representation of phenotype algorithms, and use PheKB - developed within the eMERGE Consortia - to provide a collaborative authoring, accessing and sharing of QDM-based phenotyping algorithms. These algorithms conform to Meaningful Use standards, and hence are amenable to implementation and execution within multiple EHR settings. This is enabled by using advanced workflow and rules-based technology - KNIME - by automatically translating and executing QDM-based phenotype definitions for execution within i2b2 installations as well as institution-specific data warehouses.

System status. At the time of submitting this proposal, an alpha version of PheMA is available at \url{http://projectphema.org}. The source code is released under Apache 2 licensing, and available at \url{https://github.com/PheMA}.

References

ePAD: Leveraging image data in learning healthcare systems

Daniel L. Rubin, MD, MS,1,2 Dev Gude,1 Cavit Altindag,1 Hakan Bulu, PhD,1 and Debra Willrett, MS1
1Department of Radiology and 2Medicine (Biomedical Informatics Research), Stanford, CA, USA

Abstract
As biomedical informatics efforts are undertaken to build the learning health system, there is a need to include the information provided by medical imaging in these efforts, since imaging provides detailed information about the disease phenotype for diagnosis and its response to treatment. However, at present, radiology images are not leveraged in many healthcare applications (other than viewing the raw images) because the disease phenotype information they contain is unstructured and not directly machine-accessible. We developed the electronic Physician Annotation Device (ePAD), a freely-available Web-based platform for capturing and storing the phenotypic information contained in radiological images (quantitative and semantic image features) in an explicit, standardized, and machine-accessible format that is interoperable with medical standards such as DICOM and HL7. The ePAD platform is extensible, permitting the community to extend its capabilities with respect to extracting and computing image features, as well as enabling developers to build applications that leverage the information in images in combination with other clinical data. ePAD is being used at several institutions internationally as well as in national resources such as The Cancer Genome Atlas (TCGA) project of the NIH to enable a coordinated national collection of minable radiological image data.

System Description
We will demonstrate ePAD (http://epad.stanford.edu), a freely-available Web-based system to enable researchers who view radiological images to annotate them to make the semantic contents explicit and machine-accessible. It also computes quantitative image-based features of disease, creating a comprehensive representation of the semantic and quantitative contents of images (“image metadata”) as “image annotations.” ePAD runs in a Web browser as a rich Web application, with a virtual machine back end that stores the images and image metadata. ePAD is not FDA approved and is presently intended for use in research, though its principles of standardized capture of image metadata can be incorporated into commercial image viewing workstations and clinical practice in the future.

The ePAD tool displays images and collects information from the user about the “image phenotypes” of disease (semantic features) using structured reporting templates superimposed on the images. Users can draw regions of interest and invoke a variety of tools (extensible via a plugin mechanism) to compute quantitative image features. All of semantic and quantitative image data are collected seamlessly as part of the routine image viewing workflow, and all image metadata are saved in the standardized Annotation and Image Markup (AIM) format developed by the NCI. Users can view and make image annotations very quickly via a display which shows all the abnormalities that were annotated previously; this enables users to view images and annotations with a single click of the mouse. ePAD also enforces minimum information requirements in image annotations, e.g., lesion name, measurements, a lesion type, and the anatomic location of the lesion. Semantic terms in ePAD are matched to ontologies.

We will demonstrate using ePAD to view and annotate images from cancer patients, and show new developments that enable user to build applications on the ePAD platform as plugins that leverage the image data to (1) reveal image-based treatment response for oncologist decision support, to (2) retrieve images containing similar-appearing abnormalities for radiologist decision support, and (3) compute novel quantitative imaging biomarkers and integrate them into clinical trials to improve assessment of cancer treatments. We will also discuss interoperability with DICOM and HL7 standards. Technical details about building plugins and applications with ePAD will be presented as well as how to become involved in submitting feedback and future plans for ePAD development.

ePAD is assuming a key role in national collaborative projects to build databases of minable image metadata for discovery and to incorporate image data into learning healthcare systems. The TCGA project is using ePAD to coordinate multiple centers who are collaboratively annotating radiology images of brain cancers so they can discover semantic and quantitative features of imaging phenotypes that identify disease subtypes correlated with survival. Multiple centers internationally are developing plugins to the ePAD platform to extend its capabilities and to create applications that leverage annotated image data for a variety of clinical use cases, such as content based image retrieval, radiology/pathology correlation, and linking image features to molecular data for radiogenomics discovery. A recent NCI-funded effort is using ePAD in a national effort to collect image annotations of radiology images to establish a ground truth in challenge competitions to optimize image segmentation, compare quantitative image features, and identify new image features that predict cancer treatment response and other clinical outcomes.
The SDIDS System for Integrating Global Health Surveillance Data: An Example Application to Malaria Surveillance in Uganda

Kate Zinszer, PhD1, Anya Okhmatovskaia, PhD1, Arash Shaban-Nejad, PhD1, Lauren Carroll, MPH2, Neil Abernethy, PhD3, David L. Buckeridge, MD PhD1
1McGill University, Montreal, QC, Canada; 2University of Washington, Seattle, WA

Abstract

Data that could be used for global disease surveillance are divided across diseases, countries, and organizations. The integration of this data requires substantial effort and malaria surveillance is an example of how data fragmentation can hinder evidence-based decision-making. In order to achieve allocative efficiency in malaria control programming, access to timely and accurate malaria surveillance data across a variety sources is necessary. The Scalable Data Integration for Disease Surveillance (SDIDS) is a software application designed to enable the integration and analysis of data across multiple scales to support global health decision-making. We present a prototype of SDIDS to demonstrate how it can be used to integrate malaria surveillance data collected by multiple organizations in Uganda. SDIDS is a web-based platform that automates the integration of heterogeneous data from multiple sources, and supports visualization, analysis, and sharing of these data. Specifically, it contains semantic integration of disparate data sources, automatic computations of health indicators, and an open interface for requesting standardized data and analysis.

Introduction

The growing volume of data generated by systems and devices around the world presents a tremendous opportunity for global health surveillance. These data, however, are fragmented, or “siloed” in many ways, across diseases, countries, governmental and non-governmental organizations, and clinical institutions. This fragmentation poses a barrier to analyses that could benefit from using multiple data sources as the data must be integrated manually for each analysis, requiring a considerable amount of effort. Malaria surveillance, including the monitoring and evaluation of malaria control activities, is an example of how data fragmentation can hinder analyses. Matching effective interventions to specific factors in each area requires epidemiological analysis of data integrated across multiple sources. This type of data is generally not available to malaria control and elimination activities due to the challenges in identifying, accessing, and integrating data hosted within the wide range of organizations contributing to malaria control efforts. SDIDS (Scalable Data Integration for Disease Surveillance) is a software application designed to enable the integration and analysis of data across multiple scales to support global health decision-making. In this demonstration, we present a prototype of SDIDS and show how it can be used to integrate malaria surveillance data collected by multiple organizations in Uganda.

System Description

SDIDS is a web-based, ontology-driven software platform that automates the integration of heterogeneous data from multiple sources, and supports visualization, analysis, and sharing of these data. SDIDS contains the following features:

• **Semantic integration of disparate data sources.** A uniform data representation framework encoded in an ontology is used within SDIDS to integrate existing data sources. The current prototype integrates data from over eleven administrative and clinical sources describing factors such as clinical care, interventions, and demography.

• **Access to a variety of health indicators.** The system automatically computes numerous health indicators from malaria-related data and provides multiple stratification options. The indicators are organized in a meaningful way along multiple axes using a domain ontology, which encodes knowledge about malaria.

• **An open interface for requesting standardized data and analysis.** External applications can connect directly to SDIDS to request data for further processing or to request the results of analyses applied to the integrated data. Three such applications have been developed to demonstrate the functionality of this interface in SDIDS.

A central characteristic of SDIDS is its ability to scale-up and integrate data from other geographical regions and for other priority diseases. This scalability means that a wide range of data sources can be mapped once to SDIDS and then accessed and analyzed repeatedly by a wide range of global health users and applications.

References